Table of Contents

Research on House Price Prediction Based on Multi-Dimensional Data Fusion

Yang Yonghui
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: yangyh26@qq.com

***Abstract*—The price of commercial housing is related to the process of urbanization in China and the living standard of residents, so the prediction of the price of commercial housing is very important. A major difficulty in predicting regression problems is how to handle different attribute types and fuse them. This paper proposes a house price prediction model based on multi-dimensional data fusion and a fully connected neural network. The model building steps are: First, normalize the data involved in the sample; then, interpolate the normalized data to increase the data density; subsequently, the normalized sample data is converted into a pixel matrix; finally, a fully connected neural network model is established from the pixel matrix to the price of the commercial house. After the neural network model has been established, the price of house can be obtained by entering the attributes of the house into the neural network model.**

Keywords-Multi-Dimensional Data Fusion; Fully Connected Neural Network Model; House Price Prediction

I. INTRODUCTION

Urbanization[1], also known as urbanization and urbanization, refers to the process of population gathering towards cities, the expansion of cities, and the series of economic and social changes that result from it. The essence is the changes in economic, social, and spatial structures. Modernization is the core proposition of China's modernization process and sustained economic growth. In recent years, with the further progress of China's urbanization process, more and more young people have begun to enter second-tier, third tier and even first-tier cities. A major factor affecting young people's entry into big cities is the price of local commercial housing. In other words, a major factor affecting China's urbanization process is the price of urban house. This shows that it is necessary to forecast house prices. The attributes that affect house prices are transaction date, house age, distance from the subway station, the number of convenience stores in the walking circle, the dimension of the house, and the longitude of the house. This paper will build a data fusion model. The input information of this model is the seven factors that affect house prices, and the output information is the price of commercial housing. After the data fusion model has been established, only the attributes that affect house prices are entered into the data fusion model, and the price of the commercial house can be obtained.

A. Research Background and Significance

With the development of China's economy, people's living standards have gradually improved, and economic development has made people have a higher pursuit of living places. According to data from the National Bureau of Statistics[2]: from January to December 2018, the investment in real estate development nationwide was 12,266.4 billion yuan, an increase of 9.5% over the previous year, and the growth rate was 0.2 percentage points lower than the January-November period, an increase from the same period of the previous year. 2.5 percentage points. Among them, residential investment was 8,529.2 billion yuan, an increase of 13.4%, a 0.2 percentage point drop from January to November, and an increase of 4 percentage points from the previous year. The proportion of residential investment in real estate development investment was 70.8%. With the increase in housing sales, housing prices have also increased. According to relevant data, China's housing prices have at least doubled from 2015. With the increase of house prices, people pay more attention to the prediction of house prices. This paper will build a data fusion model. The input information of this model is six attributes that affect house prices: transaction date, house age, distance from the subway station, the number of convenience stores in the walking circle, the dimension of the house, and the longitude of the house; the output is the price of the commercial house. After the data

DOI: 10.21307/ijanmc-2020-001

fusion model has been established, only the six attributes that affect house prices are entered into the data fusion model, and the price of the commercial house can be obtained. The research of house price prediction based on multi-dimensional data fusion can provide reference for China's house price prediction and further promote the development of urbanization in China.

B. Data sources

The data in this paper comes from the Boston house price data provided by Kaggle, and the amount of data selected is relatively small. The data set contains 404 training samples and 102 test samples, for a total of 506 sample data. There are 6 attributes that affect house prices in house price forecasts. In the problem of house price prediction, the attributes that affect house prices are: transaction dateX_1, house ageX_2, distance from the subway stationX_3, the number of convenience stores in the walking circleX_4, the dimension of the houseX_5, and the longitude of the houseX_5; Dependent variable is house priceY.

II. KEY TECHNOLOGY

A. Research methods for regression problems

House price forecasting is a forecasting problem, and forecasting problems are regression analysis. This section aims to state the research methods of regression analysis. Regression analysis[3] is a method of statistically analyzing data. The purpose is to understand whether two or more variables are related, the direction, and strength of the correlation and establish a mathematical model to observe specific variables to predict the variables of interest to researchers. The roles in regression analysis are independent and dependent variables: the independent variable is a variable that actively changes, for example, several factors that affect house prices in this paper are independent variables; the dependent variable is a passively generated due to changes in independent variables, such as housing prices in this paper, are a dependent variable. Regression analysis can also be understood as a method for analyzing the relationship between independent and dependent variables. The regression analysis methods are linear regression, logistic regression, and polynomial stepwise regression.

Linear regression is a linear equation established between the independent variable and the dependent variable. This is the most well-known regression model. In this type of model, the independent variable may be discrete or continuous; the dependent variable must be continuous, and the nature of linear regression is linear. Logistic regression is a logistic equation built from independent variables to dependent variables. This is a regression model used to calculate the success or failure of an event. In this type of model, the independent variable may be discrete or continuous; the dependent variable must be in the interval $[0,1]$. Polynomial regression is a polynomial equation established between the independent variable and the dependent variable. This is a polynomial regression model commonly used in the field of deep learning. Under this model, a low polynomial degree leads to underfitting, and a high polynomial degree leads to overfitting. When dealing with multiple independent variables, stepwise regression is needed[4]. Standard stepwise regression does two things, adding or removing independent variables at each step. In this technique, the selection of independent variables is done by means of an automated process, which does not involve manual intervention.

B. Research methods for data fusion

Data fusion[5] is a technology that fuses attribute values from different attributes. Fusion of multiple attributes will get better performance results than a single attribute. Data fusion is widely used in multidisciplinary and multi-scenario integration fields. For example, you can monitor the patient's physiological and psychological information through different hardware devices, and finally obtain the patient's physical condition through data fusion. There are many similar examples. There are also many difficulties in data fusion. The first is how to deal with different attributes, and the second is how to fuse the data.

There are many difficulties in data fusion design. The first is how to handle different attribute types, and the second is how to fuse attributes. This thesis will detail the processing method of the attribute type in the "Handling of attribute types" Section and the data fusion method in the "Data Fusion" Section.

C. Handling of attribute types

The attribute type refers to the data type of the attribute. The attribute types are: Large_Attributes, Small_Attributes, Intermediate_Attributes, and Interval_Attributes[6].

1) Large_Attributes

The Large_Attributes are the larger the independent variable, the larger the dependent variable, that is, the independent variable will have a positive benefit on the dependent variable, in other words, there is a positive correlation between the dependent variable and the independent variable. The processing method for very large attributes is shown in (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Among them, x_{max} is the maximum value of the attribute value; x_{min} is the minimum value of the attribute value; x is the original value of the attribute value; x' is the normalized attribute value.

2) Small_Attributes

The Small_Attributes refers to: the larger the independent variable, the smaller the value of the dependent variable, that is, the independent variable will have a negative benefit on the dependent variable, in other words there is a negative correlation between the independent variable and the dependent variable. The processing method of extremely small attributes is shown in (2).

$$x' = \frac{x_{max} - x}{x_{max} - x_{min}} \quad (2)$$

Among them, x_{max} is the maximum value of the attribute value; x_{min} is the minimum value of the attribute value; x is the original value of the attribute value; x' is the normalized attribute value. After processing by the above method, the extremely small attributes have been transformed into extremely large attributes.

3) Intermediate_Attributes

Intermediate_Attributes refer to the existence of a threshold. When the independent variable is smaller than the threshold, it displays the characteristics of Large_Attributes. When the independent variable is larger than the threshold, it displays the characteristics of Small_Attributes. Specifically, when the independent variable is less than the threshold, there is a positive correlation between the independent variable and the dependent variable; when the independent variable is greater than the threshold, there is a negative correlation between the independent variable and the dependent variable. The processing method of Intermediate_Attributes is shown in (3).

$$\begin{cases} x' = \frac{x - x_{min}}{x_0 - x_{min}}, x < x_0 \\ x' = \frac{x_{max} - x}{x_{max} - x_0}, x > x_0 \end{cases} \quad (3)$$

Among them, x_{max} is the maximum value of the attribute value; x_{min} is the minimum value of the attribute value; x is the original value of the attribute value; x' is the normalized attribute value; x_0 is the threshold. After processing by the above method, the interval attribute has been transformed into Large_Attributes.

4) Enumerated_Attributes

Enumerated_Attributes means that the attribute value of the independent variable does not have real measurement characteristics, and the result of the dependent variable will be affected by the value of the independent variable, but this influence relationship is difficult to express. The processing method of Enumerated_Attributes is as follows:

Step1: List all the values of the input attributes;

Suppose the input attribute contains l attribute values: x_1、x_2、…、x_l;

Step2: Convert the attribute value to One-Hot [7] form;

Among them, x_1 is the $1st$ attribute value, so a vector with only the $1st$ position being 1 can be used instead. That is, x_1 can be expressed as: $(1 \quad 0 \quad \cdots \quad 0 \quad \cdots \quad 0 \quad \cdots \quad 0)_{1\times l}^{T}$;

Among them, x_2 is the $2nd$ attribute value, so a vector with only the $2nd$ position being 1 can be used instead. That is, x_2 can be expressed as: $(0 \quad 1 \quad \cdots \quad 0 \quad \cdots \quad 0 \quad \cdots \quad 0)_{1\times l}^{T}$; ……

Among them, x_l is the lst attribute value, so a vector with only the lst position being 1 can be used instead. That is, x_l can be expressed as: $(0 \quad 0 \quad \cdots \quad 0 \quad \cdots \quad 0 \quad \cdots \quad 1)_{1\times l}^{T}$;

So far, all values of the attribute have been expressed as One-Hot form.

D. Data Fusion

This section analyzes the problem of data fusion, that is, how to merge Large_Attributes, Small_Attributes, Interval_Attributes, and Enumerated_Attributes together. This thesis will propose a pixel-based data fusion method: first establish a pixel matrix; then use a fully connected neural network model to process the pixel matrix.

1) Create a pixel matrix

This section aims to transform multiple attributes into a pixel arrangement. Specifically, it is assumed that the sample contains m samples and each sample contain n attributes, that is,

All values for the $1st$ sample are: x_{11}, x_{12}, …, x_{1i}, …, x_{1j}, …, x_{1m};

All values for the $2nd$ sample are: x_{21}, x_{22}, …, x_{2i}, …, x_{2j}, …, x_{2m};

……

All values for the mnd sample are:x_{m1}, x_{m2}, …, x_{mi}, …, x_{mj}, …, x_{mm}.

Then, the $1st$ pixel matrix is:$(x_{11} \quad x_{12} \quad \cdots \quad x_{1i} \quad \cdots \quad x_{1j} \quad \cdots \quad x_{1m})^T$;

and the $2nd$ pixel matrix is:$(x_{21} \quad x_{22} \quad \cdots \quad x_{2i} \quad \cdots \quad x_{2j} \quad \cdots \quad x_{2m})^T$;

……

and the $2nd$ pixel matrix is:$(x_{m1} \quad x_{m2} \quad \cdots \quad x_{mi} \quad \cdots \quad x_{mj} \quad \cdots \quad x_{mm})^T$.

2) Processing pixel matrix

In "Create a pixel matrix", this article has already established the number of pixel matrices as the number of samples, and then we need to use the neural network to process the pixel matrix.

The choice of network structure: there are many neural network model structures, such as fully connected layer neural networks, convolutional neural networks, long-short-term memory networks, and Residual network. Because the application scenario in this paper is simple, it is more appropriate to choose a fully connected neural network model.

Selection of activation function: The activation function is a function that runs on the neuron and is responsible for mapping the input of the neuron to the output. The activation functions are: $Sigmoid$ function (Figure 1 $Sigmoid$), $Tanh$function (Figure 2 $Tanh$), $ReLU$ function (Figure 3 $ReLU$), $Leaky\ ReLU$ function (Figure 4 $Leaky\ ReLU$), where $Leaky\ ReLU$ is a special form of $ReLU$. Regarding the selection principle of the activation function, Andrew Ng gives the following reference scheme in "Neural Networks and Deep Learning": $Tanh$ is very common in machine learning. The activation function is generally defaulted to $Tanh$. $Leaky\ ReLU$ is generally better than $ReLU$, but the scope of use of $ReLU$ Wider; the activation function used in the output layer of the binary classification problem is $Sigmoid$, and $Sigmoid$ was rarely used in other cases; $Tanh$ is almost always better than $Sigmoid$. $Tanh$ and $Sigmoid$ have a disadvantage that when the independent variable is large, the slope is small. The gradient descent method is limited; except for the output layer, linear activation functions are rarely used; neural network models use activation functions, which will lead to the final result being a linear combination of input feature vectors.

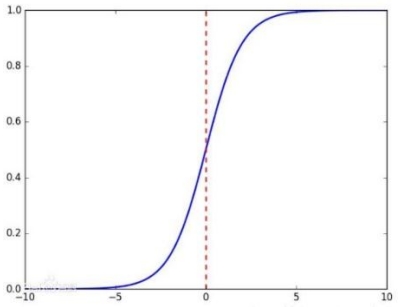


Figure 1. Sigmoid

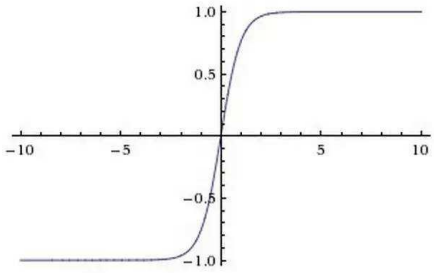


Figure 2. Tanh

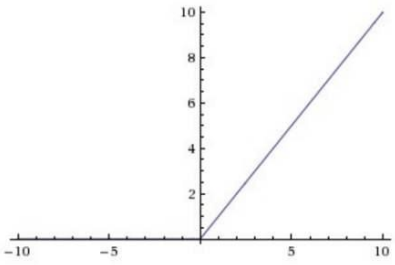


Figure 3. ReLU

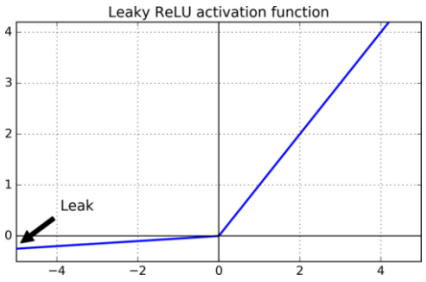


Figure 4. Leaky ReLU

III. NORMALIZATION OF ATTRIBUTES

This part needs to normalize the attributes involved in the data set: first analyze the data type of the attributes by "Attribute Analysis"; then normalize the attributes by "Normalization".

A. Attribute Analysis

As mentioned in "Data Sources", the data in this paper is derived from Boston house price data provided by Kaggle, and the amount of data selected this time is relatively small. The data set contains 404 training samples and 102 test samples, for a total of 506 sample data. In the problem of house price prediction, there are 6 attributes that affect house prices: transaction date X_1; house age X_2; distance from the subway station X_3; the number of convenience stores in the walking circle X_4; the dimension of the house X_5; the longitude of the house X_6;. dependent variable: house price Y.

Transaction date X_1 is a time variable; the house age X_2 is a Small_Attributes; distance from the subway station X_3 is a Small_Attributes; the number of convenience stores in the walking circle X_4 is a Large_Attributes; the dimension of the house X_5 and the longitude of the house X_6 are an Enumerated_Attributes.

B. Normalization

Transaction date X_1 is a time variable; the house age X_2 is a Small_Attributes; distance from the subway station X_3 is a Small_Attributes; the number of convenience stores in the walking circle X_4 is a Large_Attributes; the dimension of the house X_5 and the longitude of the house X_6 are an Enumerated_Attributes.

IV. Data Fusion

In this part, the normalized data in "Normalization of attributes" needs to be fused: first, the pixel matrix is established by "Building a Pixel Matrix"; then the fully connected neural network model is established by "Building a Neural Network Model".

A. Building a Pixel Matrix

A pixel matrix can be established by "Data Fusion". As described in "Data sources", the data in this paper is derived from Boston house price data provided by Kaggle. The amount of data selected is small. The data set contains 404 training samples and 102 test samples, for a total of 506 sample data. Then there are:

All values for the $1st$ sample:$x_{11}, x_{12}, \ldots, x_{17}$;

All values for the 2nd sample:$x_{21}, x_{22}, \ldots, x_{27}$;

……

All values for the 506st sample:$x_{506\_1}, x_{506\_2}, \ldots, x_{506\_7}$.

B. Building a Neural Network Model

The paper will eventually build a neural network model of house attributes to house prices: where the input attributes are house attributes: transaction date X_1; house age X_2; distance from the subway station X_3; the number of convenience stores in the walking circle X_4; the dimension of the house X_5; the longitude of the house X_6;output information is house price Y.

Step1: Design the network structure

Through the analysis of "Data Fusion", this paper will build a fully connected neural network model. The network model structure is shown in (Figure 5 Network structure): The input layer of the network structure contains 7 input nodes; the network structure contains 5 hidden layers, each of which contains 4 nodes; the output layer of the network structure contains 1 output node; all activation functions use $ReLU$ function; Training period: 50000; Target accuracy is: 10^{-5}; Learning rate: 0.01

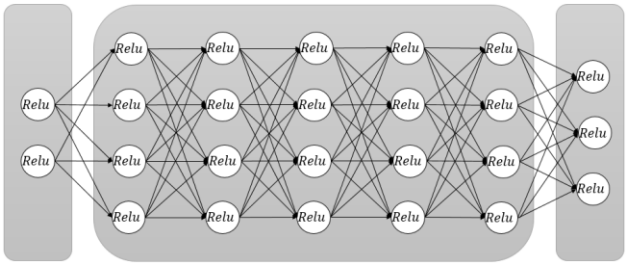


Figure 5. Network structure

Step2: Selection of training tools

There are many ways to train neural networks, such as Tensorflow, Caffe, MXNet, Torch, Theano in python, and nntool in Matlab. nntool is a network model training tool that is easy to deploy and simple in the environment. In this paper, the neural network model shown in (Figure 5 Network structure) is trained by nntool (Figure 6 nntool).

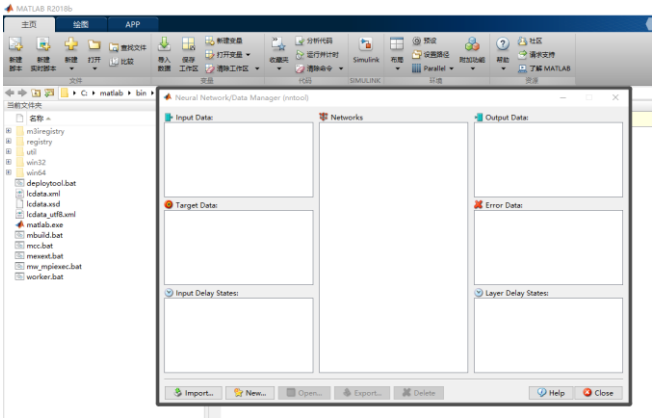


Figure 6. Nntool

Step3: Code design
See Appendix

Step4: Training process
In the process of neural network training using Matlab, part of the training process is shown in (Figure7 Training process). Among them, Performance is shown in (Figure 8 Performance); Training State is shown in (Figure9 Training State); Regression is shown in (Figure10 Regression).

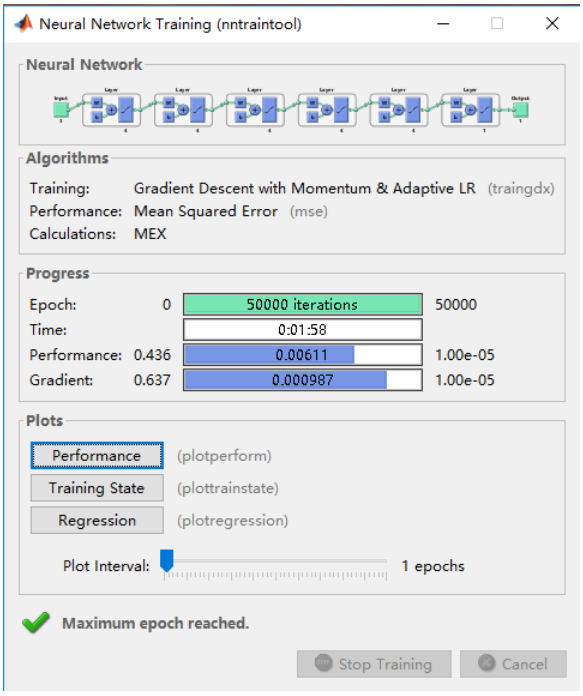


Figure 7. Training process

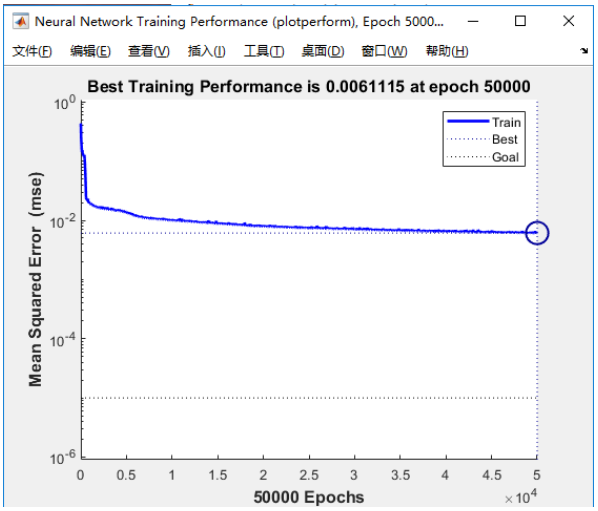


Figure 8. Performance

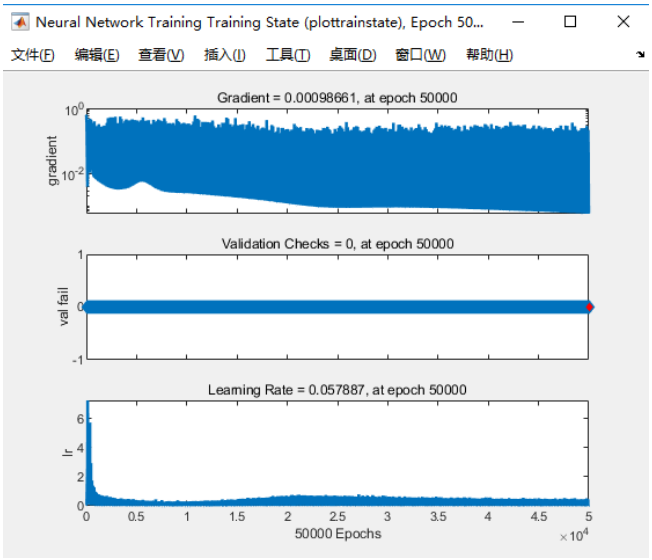


Figure 9. Training State

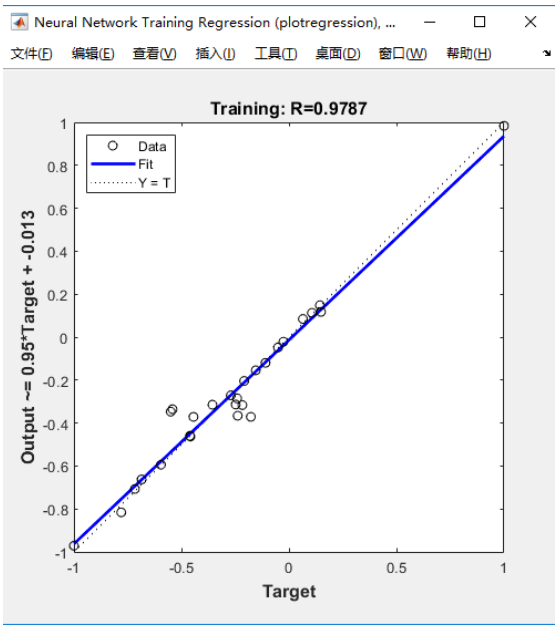


Figure 10. Regression

Step5: Results
The results of the neural network model include two parts: one is the partial result display, as shown in (Figure11 Result); the other is the error proportion chart, as shown in (Figure12 error_raph). As can be seen from the (Figure10 Regression), the accuracy of the network model is 97.87%.

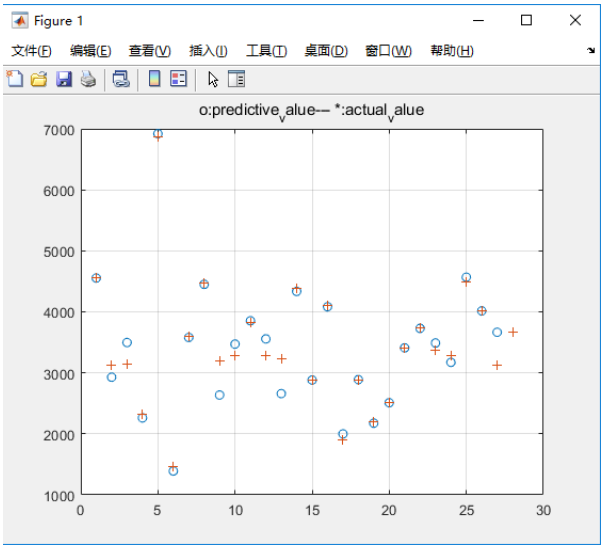


Figure 11. Result

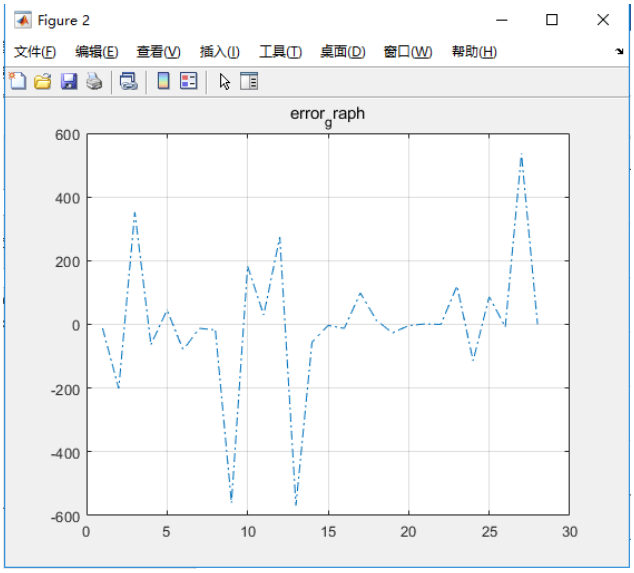


Figure 12. Error_raph

V. SUMMARY

This paper finally established a neural network model from house attributes to house prices: where the input attributes are commodity house attributes: transaction date X_1; house age X_2; distance from the subway station X_3; the number of convenience stores in the walking circle X_4; the dimension of the house X_5; the longitude of the house X_6;output information is house price Y.After the neural network model has been established, Enter the six attributes of the commercial house into this neural network model, and you can get the corresponding house price. The accuracy of the network model is 97.87%.

VI. APPENDIX

```
[pn,minp,maxp,tn,mint,maxt]=premnmx(p,t);
NodeNum1 =4;
NodeNum2=4;
NodeNum3=4;
NodeNum4=4;
NodeNum5=4;
TypeNum = 1;
TF1 = 'tansig';
TF2 = 'tansig';
TF3 = 'tansig';
TF4 = 'tansig';
TF5 = 'tansig';
TF6 = 'tansig';
net=newff(minmax(pn),[NodeNum1,NodeNum2,NodeNum3,NodeNum4,NodeNum5,TypeNum],{TF1 TF2 TF3 TF4 TF5 TF6},'traingdx');
%traingdm
net.trainParam.show=50;
net.trainParam.epochs=50000;
net.trainParam.goal=1e-5;
net.trainParam.lr=0.01;
net=train(net,pn,tn);
p2n=tramnmx(ptest,minp,maxp);
an=sim(net,p2n);
[a]=postmnmx(an,mint,maxt)
plot(1:length(t),t,'o',1:length(t)+1,a,'+');
title('o:predictive_value--- *:actual_value')
grid on
m=length(a);
t1=[t,a(m)];
error=t1-a;
figure
plot(1:length(error),error,'-.')
title('error_graph')
grid on
```

REFERENCES

[1] Lee W C, Cheong T S, Wu Y. The Impacts of Financial Development, Urbanization, and Globalization on Income Inequality: A Regression-Based Decomposition Approach [J]. SSRN Electronic Journal, 2017.

[2] Tan Paul. House prices have been stagnant [J]. Journalist observation, 2019 (4).

[3] Gogtay N J, Deshpande S P, Thatte U M. Principles of Regression Analysis [J]. The Journal of the Association of Physicians of India, 2017, 65(4):48-52.

[4] Gooch J W. Stepwise Regression [J]. Encyclopedic Dictionary of Polymers, 2011.

[5] Bleiholder J, Naumann F. Data fusion [J]. ACM Computing Surveys, 2008, 41(1):1-41.

[6] Han Zhonggeng. Mathematical model for comprehensive evaluation and prediction of Yangtze River water quality [J]. Journal of Engineering Mathematics (7): 69-79.

[7] Shuntaro Okada, Masayuki Ohzeki, Shinichiro Taguchi. Efficient partition of integer optimization problems with one-hot encoding[J]. Scientific Reports, 2019, 9(1).

[8] Wang Zhaoqing, Lu Xiaoyang. A Macro Element Method for Solving Potential Problems Based on Mean Value Interpolation [J]. Journal on Numerical Methods and Computer Applications (3): 21-29.

[9] Hershey S, Chaudhuri S, Ellis D P W, et al. CNN architectures for large-scale audio classification[C]// 2017.

Research on Real-Name Routing and Trusted Connection Based on IPV9 and CPK-card

Xie Jianping
[1.]Chinese Decimal Network Working Group, Shanghai, China
[2.]Shanghai Decimal System Network Information Technology Ltd. Shanghai, China
E-mail: 13386036170@189.cn

Nan Xianghao
[1.]Chinese Decimal Network Working Group, Shanghai, China
[2.]Shanghai Decimal System Network Information Technology Ltd. Shanghai, China
E-mail: nanxh2001@163. com

AliAbZoraghchin
Department of Computer Engineering Allame Mohaddes Noori Institute of Higher Education Mazandaran, Iran
E-mail: aliab@yahoo.com

***Abstract*—Router is the basic component of the Internet. In this scheme, identification technology is used for the first time in router design to provide address authenticity proof to prevent illegal access. Provide proof of connection freshness to prevent reissue attacks; the first use of software identification technology, to provide the credibility of the router operating environment, to prevent Trojan and other malicious software intrusion. The design also provides densification function to ensure privacy. This is a key security requirement for the next generation of Internet protocols. This design method will be combined with the new addressing technology of geographical location addressing to construct the next generation of Internet routers. The technology is also used in the design of new switches in telecommunications networks.**

Keyword-IPV9; CPK-card; Real-Name Routing; Trusted Connection

I. INTRODUCTION

Routers work in the network layer of the OSI seven-layer protocol. Their main function is to connect the network with the network and forward packets between the networks. Routers have become the most important network equipment, so the research on the new generation of routers will become the core technology of the next generation of Internet research. Due to the IPv4, IPv6 protocols that used in the Internet, the new requirements for trusted Cyber Security connections are not met. The TCP/IP protocol has no security concerns, it does not provide address authentication, does not prevent unauthorized access, and does not protect against DOS attacks.

At present, all kinds of malicious software and spam information are rampant on the Internet, which seriously pollutes the use environment of the Internet and directly affects the survival of the Internet. As a result, all countries over the world have developed a new generation of green Internet research. In 2008 the European Union's 65 scientific institutions jointly issued the brad declaration, calling for a new generation of the Internet. The European Union has raised 9.1 billion Euros to support future Internet research and development. The U.S governments have also successive proposed identity authentication and Addressing system as major scientific research tasks, and emphasized international cooperation. ISO, the international standards body, put forward its plan for the future network in 2007.

In 2007, Chinese researchers Xie Jianping proposed the IPv9 geographical location addressing method, which solved the problem of combining IP address with geographical location. Later, South Korea also proposed the idea of geographical location addressing, and becoming the second country to propose a new method of addressing. CPK (Combined Public Key) identity authentication technology is mature and can be used in Internet protocol to realize trusted connection.

So far, China had already had the technology foundation that research and development next generation router and future network protocol.

II. REQUIREMENTS FOR TRUSTED CONNECTIONS

In order to realize the trusted connection between routers and users, the user name (Pc1) and route address (Alfa) are identified for identity authentication. Among routers, mutual authentication is made with IP address as identity, and mutual authentication is made with user name as identity between users. Suppose that Pc1ID is the user name of a client and AlfaID is the IP address of a router. Assume that AlfaID="china-beijing-haidian-peking university" and BetaID="china-beijing-haidian-tsinghua university".

Now assume the starting address is AlfaID and the destination address is BetaID, and the connection process is shown in figure 1 (dotted lines indicate that CPK-card is used and the original address is identified)..

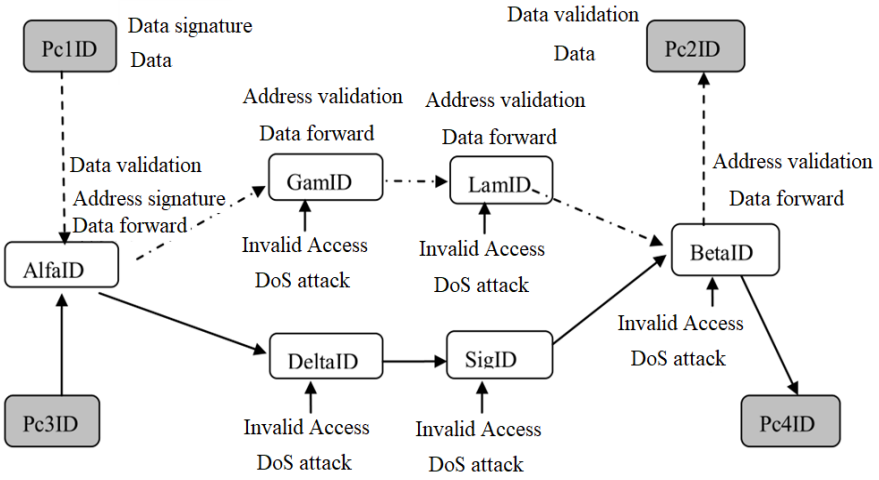


Figure 1. Data workflow diagram

The IP packet of the original router passes through multiple routing routers and finally arrives at the destination router. Illegal access is easy to access in the intermediate routing router. It can be seen from the working principle of the above router that previous routers only pay attention to the routing of the next hop and do not care where the packet comes from. Therefore, if we do not solve the origin address verification, we cannot overcome the illegal access.

Some people try to solve the problem of illegal access by means of encryption, but under the condition of public key system, this is futile. For example, Beta is the receiving party, and its public key is public, so anyone can encrypt Beta, so Beta still doesn't know who the sender is.

In order to achieve the trusted connection, the router must meet the following four conditions:

1) The primary IP address must be given to send proof of address; it can be verified by any one place;

2) All routing routers verify the original address, and reject forward if there is any discrepancy;

3) It can prevent illegal access and resist DOS attacks.

4) The internal computing environment of the router is reliable.

III. IPV9 CONNECTION ENVIRONMENT

The connection policy is as follows.

Path 1 (all using IPV9 protocol and CPK-card)

1) Pc1ID signs the data with pc1 and delivers the signed data to the router AlfaID.

2) AlfaID signs the time with alfa and forwards it to the next router, which verifies the signature of the original address. If the verification passes, the data is forwarded to the next router.

3) GamID, LamID, BetaID and other routing operations are the same as above.

4) The BetaID route forwards the data to the receiving user Pc2ID.

Path 2 (Client adopts IPV9 protocol, but does not use CPK-card) :

1) Pc3ID does not use CPK-card but sends data to Pc4ID via routing AlfaID via PT converted to IPV9 protocol.

2) The AlfaID route obtains the packet source address as the public key and verifies the correctness of the source.

Path 3 (The client does not use IPV9 protocol and CPK-card):

1) Pc3ID does not use CPK-card and USES IPV4/IPV6 protocol to route data to Pc4ID via AlfaID.

2) Route through the DeltaID and SigID routes to the BetaID route and forward the data to PC4ID.

Path 4 (The client adopts IPV9 protocol and USES CPK-card, but the middle V9 route does not use CPK-card):

1) Pc1ID using the local address as the public key to sign the data and sends data to Pc2ID via route AlfaID.

2) The AlfaID route takes the source address of the packet as the public key and verifies the correctness of the source. After verification of the source address, remove the original signature and use the local address as the public key signature. After the signature, forward the normal routing data.

3) Instead of using CPK-card, GamID obtains the source address of the packet as the public key and verifies the correctness of the source. If the address is not legitimate, the data is discarded and the normal routing data is forwarded.

4) LamID, BetaID and other routing operations are the same as above.

5) The BetaID route forwards the data to the destination Pc2ID.

The IPV9 v4/v6 compatible data forwarding process is shown in figure 2.

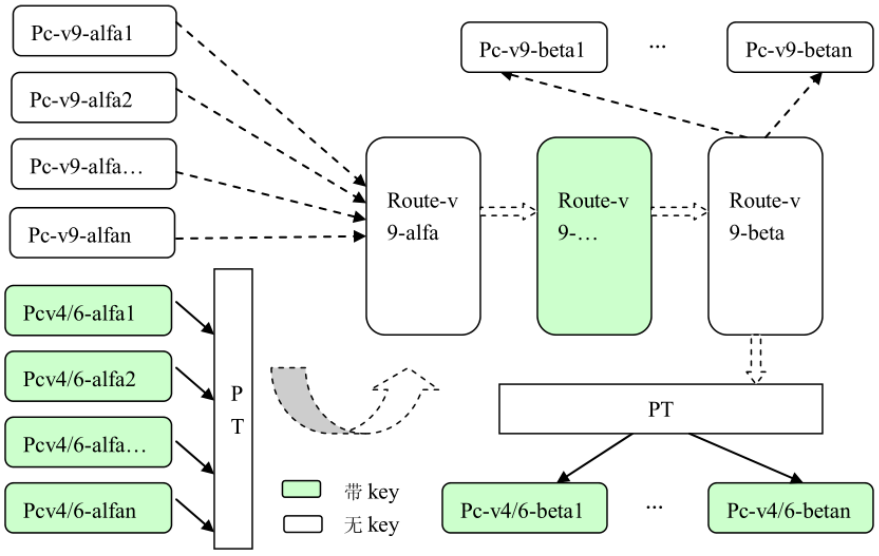


Figure 2. Data forward process diagram

IV. IPV9 AUTHENTICATION FUNCTION

A. *CPK-Card*

CPK is a public key-based cryptography system that takes the identity of any entity directly as the public key, while the private key is distributed in the form of ID-card. Now, for example PC1, ALFA (uppercase) and so on respectively represent their public keys, pc1, alfa (lowercase) and so on respectively represent their private keys. If it has been insert a CPK-card defined as AlfaID on any router, the router becomes the identified router as AlfaID. Similarly, any router inserts a CPK-card defined as BetaID, and the router becomes the identified router as BetaID.

The router is configured with CPK-card, which has the functions of digital signature and key exchange. The contents of CPK-card are as follows: let the router's IP address be alfa (alfa may be the real name of China. Beijing. Haidian. Peking University etc. and it can be changed into machine executable code after the unified name translation). ID-card format and size is as table 1.

B. Original address identification

Suppose the original place is AlfaID, the next router is GammaID, AlfaID sends data, Mas1 AlfaID→GammaID:{Alfa, sign1, Beta, time data, checksum}

Where, sign1 is the signature of the original AlfaID address, that is sign1= SIGalfa (time), BetaID is the destination address, SIG is the signature function, and alfa is the private key of the signature, provided by CPK-card. Where data is data from the application layer, data may be plaintext or ciphertext. The router's job is to pass the data to the next router.

TABLE I. ID-CARD FORMAT AND SIZE

1	Z1: Validate parameter	16B	EPWD(R1)=Z1
2	Z2: Validate parameter	16B	ER1(R1) ⊕ R1=Z2
3	Identify definition	25B	alfa
5	Private key 1	32B	ER1(csk1)=Y1
6	Private key 2	32B	ER1(csk2)=Y2
10	Issue unit	25B	KMC
11	Signature of issue unit	48B	SIGkmc （MAC）

GammaID verifies the signature of original address: SIG -1ALFA(time)=sign1'

Where SIG-1 is the validation function and ALFA is the public key. If sign1=sign1', allow this connection, forward Msg1, and audit. Identify replay attacks against time.

V. IPV9 ENCRYPTION FUNCTION

The structure of data is defined as follows: Data={Pc1ID,Pc2ID,data, mac}, Where Pc1ID is the sender and Pc2ID is the receiver.

When the data is in plaintext: Data={ Pc1ID, Pc2ID, clear-text, mac};

When the data is in ciphertext: Data={Pc1ID, Pc2ID, coded-key, coded-data, mac};

If the encryption and decryption function is provided by the router, and Alfa encryption and Beta decryption are set, then data encryption can only be done in a non-online way.

If the router is responsible for encryption and decryption, and this data is encrypted data, coded key and coded data need to be interpreted and a series of steps shall be performed:

1) Generate random number R, AlfaID calculation key: key=R × (G); where G is the base point of the elliptic curve; key will be used to encrypt the data;

2) 2) Calculate the sending key: R(BETA)=coded-key, where BETA is the public key of BetaID and coded-key is sent to BetaID.

3) 3) Encrypt data: Ekey (data) =cipher-text;

Send ciphertext cipher=text and coded-key to BetaID.

BetaID receives a signal from the AlfaID and automatically enters the decryption process.

- BetaID computes the inverse of the private key: beta-1 ;
- BetaID calculates the session key: beta-1(coded-key)=key;
- Data decryption：Dkey(cipher-text)= data.

VI. IPV9 PACKET HEADER FORMAT AND ENCODING FORMAT

A. Packet header format

The new features require the development of a new IP header format that includes at least the original address, the start address identifier, the destination address, the data, and the checksum. Data encryption only affects the data format, not the IP packet header format.

Version	**Category**	**Flow label**	**Payload Length**	**Next Header**	**Hop limit**
Source address					
Destination address					
time					
Identification code (signature) 40BYTE					

B. IPV9 coding format

The encoding format of IPV9 is shown in the following table 2.

TABLE II. ENCODING FORMAT OF IPV9

segment	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Head segment	Address area		Entity code	Vendor code	Product code	Product class code	
	Country code	District code				Year code	Single product code
2	4	6	4bit	14bit	20bit	8bit	199bit
		362300	3211	123345678912345	12345678912345678912	20090317	32564328

When the industrial standard "business RFID label data format" is adopted, the enterprise product coding data format is as follows:

1) The basic data format of enterprise products is as follows:

12345678912345-12345678912345678912-20090317-32564328

2) When data exchange is used between management departments, the format of enterprise product data is:

3221-12345678912345-12345678912345678912-20090317-32564328

3) When data is exchanged between regions and management departments, the format of enterprise product data is:

362300-3221-12345678912345-12345678912345678912-20090317-32564328

4) When data are exchanged between countries:

a) When exchanging with the itu-t E164 data system, the data format is:

00-8600-36230 — 3221-12345678912345-12345678912345678912-20090317-32564328.

b) When exchanging with ISO's object identifier data system, the data format is:

01-8600-362300 — 3221-12345678912345-12345678912345678912-20090317-32564328.

c) When exchanging with the object identifier data system of ISO/ITU, the data format is:

02-8600-362300 — 3221-12345678912345-12345678912345678912-20090317-32564328.

C. Enterprise product IPV9 address format

1) The basic IPV9 address format of enterprise products is:

12345678912345]12345678912345678912]20090317]32564328.

2) When data exchange is used between management departments, the IPV9 format of enterprise product data is:

3221]12345678912345]12345678912345678912]20090317]32564328.

3) When data is exchanged between regions and management departments, the IPV9 format of enterprise product data is:

362300]3221]12345678912345]12345678912345678912]20090317]32564328

4) When data are exchanged between countries, the IPV9 format is:

a) When exchanging with the ITU-T, E164 data system, the IPV9 data format is:

00]8600]362300]3221]12345678912345]12345678912345678912]20090317]32564328

b) When exchanging with ISO's object identifier data system, the IPV9 data format is:

01]8600]362300]3221]12345678912345]12345678912345678912]20090317]32564328

c) When exchanging with the object identifier data system of ISO/ITU-T, the IPV9 data format is:

02]8600]362300]3221]12345678912345]12345678912345678912]20090317]32564328

VII. IPV9 TRUSTED COMPUTING

In order to ensure the credibility of the operation of the router, all the execution code in the router must be certified by the manufacturer (level 1 certification), that is, the manufacturer sign on the appearance of all the execution code. Each router has an authentication function (provided by CPK-card).

A. Proof of software code

The manufacturer has a CPK-card, which can carry out manufacturer signature on all system software in the router. Implementation software is divided into software identity (codeID) and software ontology (codeBD), which are signed by the manufacturer respectively:

$$SIG_{manufacturer}(codeID) = sign1$$

$$SIG_{manufacturer}(codeBD) = sign2$$

Where, SIG is the signature function, manufacturer is the private key of the manufacturer, codeID is the name of the executing code, and codeBD is the HASH value of the executing code ontology. Any executing code in the router has its own authentication codes, sign1 and sign2.

B. Identification of software code

The router inserts the CPK-card so that it has the CPK authentication function. There are two ways to verify the router: one is to uniformly verify when the router is turned on, and the code that fails to pass the verification is uniformly deleted to ensure that the router system returns to the original state; the other is that when software code is invoked, it is validated first and then executed.

Verify sign1 and sign2 respectively:

$$SIG^{-1}_{MANUFACTURER}(codeID)=sign1'$$

$$SIG^{-1}_{MANUFACTURER}(codeBD)=sign2'$$

Where MANUFACTURER is the public key, it is allowed execute if sign1=sign1'and sign2=sign2', otherwise it is rejected. In this way to ensure that the implementation of the router code is the manufacturer certification code, other code will not be executed, from the attack of viruses, Trojans.

VIII. CONCLUSION

The TCP/IP protocol does not guarantee trusted connections, so it must be modified. Based on geographical encoding and location addressing, three key techniques of trusted methods are proposed in this paper. Use address identification mechanism to prevent illegal connection; Adopt random Q&A mechanism to prevent replay attack; Software code can be identified by the mechanism, to prevent the intrusion of viruses, Trojans.

The above design method is fully applicable to the trusted connection of the physical layer. There are two kinds of physical layer: one is the physical layer defined in the seven-layer information network protocol, and the platform supporting the information network is the application program interface (API). The second is the physical layer defined in the telecommunications network, and the platform supporting the telecommunications network is the telecommunications reference point (TRP). In the information network, if the network layer can guarantee the credibility of transmission, the security of the physical layer can be replaced by the network layer. However, the physical layer of the telecom network, without modification, cannot achieve trusted connection, cannot prevent illegal access. It is modified in exactly the same way as the router.

REFERENCES

[1] Tang Xiaodan etc. Computer Operating System (third edition) [M]. Xi'an: Xidian university press, 2010.

[2] Nan Xiang-hao. CPK Combined Public Key System [J]. Information Security and Communication Confidentiality, 2013(03):39-41.

[3] Xie Jianping etc. A method of assigning addresses to network computers using the full decimal algorithm [P]. CN: ZL00135182.6, 2004.2.6.

[4] Xie Jianping etc. Method of using whole digital code to assign address for computer [P]. US: 8082365, 2011.12.

[5] Nan Xianghao. CPK Public Key System and Identification Identification. [M]. Beijing: People's Posts and Telecommunications Press, 2013.

[6] Xie Jianping, Xu Dongmei, etc. Digital domain name specification. SJ/T11271-2002, 2002.07.

[7] Information technology-Future Network- Problem statement and requirement-Part 2: Naming and addressing, ISO/IEC DTR 29181-2, 2014, 12.

[8] Wang Wenfeng, Xie Jianping, etc. Product and service digital identification format for information procession. SJ/T11603-2016, 2016. 06.

[9] Radio frequency identification tag information query service network architecture technical specification. SJ/T11606-2016, 2016. 06

Research on Enterprise Application Integration Platform Based on SOA Architecture

Liu Pingping
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1341369601@qq.com

Lu Jiaxing
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1721653661@qq.com

***Abstract*—Tobacco industry is a relatively early industry in China' information industry, so there are many problems, such as the lack of overall planning, the wide application system, but the low degree of integration of information resources, and the serious problem of "information island". In order to solve how to establish an efficient and flexible way of information interaction within the enterprise, an enterprise application integration platform based on SOA architecture is proposed. The platform takes ESB as the core, transforms enterprise information integration into a new way conforming to SOA architecture, and establishes the idea of basic data management, so as to achieve the purpose of optimizing the overall information resources of the enterprise**

Keywords-SOA Framework; Information Interaction; ESB; Information Integration

I. INTRODUCTION

SOA is widely used in the IT industry in the 21st century. SOA is service oriented architecture, Service oriented architecture. It is architecture, not a technology or a method. We can also say that SOA is an idea. In China, many enterprises begin to build enterprise integration platform based on SOA. For example, Kingdee Apusic SOA. Fmqm exotica, developed by Almaden Laboratory of IBM company, is a distributed workflow management system based on persistent message queue, which can save the execution information of workflow through persistent message queue and complete the complete independence of all nodes in the execution process

SOA architecture has three significant advantages: loose coupling, coarse granularity, and location protocol transparency. Through the encapsulation of services to achieve a comprehensive loose coupling, loose coupling can reduce the dependency between services, so that the flexibility of the service itself can be improved, and it will not be forced to adjust because of other services adjustment, thus greatly improving the reusability of services. Coarse granularity means that the interface of services defined in SOA is close to the actual user operation. Location protocol transparency means that when accessing the service defined by SOA, you do not need to know the specific location and transport protocol of the service. Even if the location and transport protocol of the service change, the client that invokes the service does not need to change. Based on the investigation and analysis of the problems of information island, high coupling degree and poor integration expansibility of each system in BJ cigarette factory based on SOA architecture, an enterprise application integration platform design scheme suitable for the actual situation of the enterprise is proposed to solve the current problems.

Therefore, through the research on the practical application of the enterprise application integration platform based on SOA architecture in cigarette enterprises, through the research on SOA architecture and ESB technology, combined with the analysis of the actual problems of information integration in cigarette manufacturing enterprises, this paper puts forward the design scheme of the enterprise application integration platform that adapts to the actual situation of enterprises.

II. THE CURRENT SITUATION

Tobacco industry is an industry with an early start of information construction in China. At present, the level of information is generally high. BJ cigarette factory, as the main cigarette manufacturing enterprise in Shaanxi Province, has many application systems after years of information construction, covering all aspects of the factory from production to management. The main information systems include manufacturing execution (MES) system and enterprise Industrial

resource planning (ERP) system, logistics system, data acquisition system in the car room, centralized control system in the silk making workshop, power and energy management system, human resource management system, enterprise card system, etc.

The more application systems there are, the problem is not only the disunity of basic data, but also the complexity of system integration. The traditional integration method generally adopts the point-to-point mode. Each system needs a special channel to integrate Chengdu. As shown in Figure 1, the integration of N application systems will generate N * (n-1) integration channels with high complexity. When new application systems need to be integrated, the complexity improvement will also be exponential.

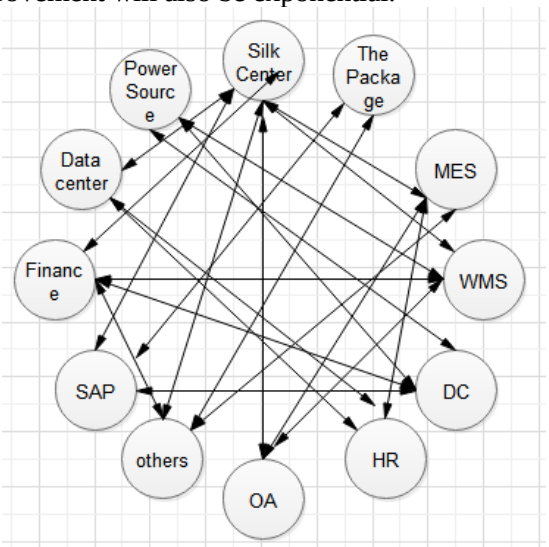


Figure 1. Integration complexity

In summary, the current informatization problems of BJ cigarette factory mainly include the following aspects:

1) There are some isolated information islands and some application systems are in information closed state due to lack of external integration means.

2) The basic data in the enterprise is scattered in different application systems, which need to be maintained separately, so it is difficult to ensure the unity of the basic data of the whole enterprise, so as to "count out one place". The lack of a unified basic data code system makes information interaction difficult.

3) Basic data depends on business system and has high coupling. At present, the basic data in the enterprise mainly depends on ERP system, and the purpose of the basic data is to provide the most basic data information for all application systems of the whole enterprise, so relying on a single application system will cause unnecessary impact on the users of other basic data.

4) Poor integration scalability. At present, the information system of the whole enterprise adopts the point-to-point integration mode. If the new application system wants to join the integration system, it needs the cooperation of each application system. The upgrading or transformation of the existing application system also needs to involve a lot of external interface changes, because of this poor scalability.

5) The lack of management and monitoring of data interaction process makes it difficult to find and deal with problems in the process of data interaction in time. Some data have high requirements for timeliness. If it can not be communicated in time, it will have a significant impact on the actual business. Therefore, effective management and monitoring measures are needed for the data interaction process.

6) The integration of point-to-point results in the aggravation of network burden, and many data interaction contents are repeated, but the data can not be reused, resulting in the waste of resources.

Through the analysis and optimization of the current problems of the enterprise, the core of which is to establish a reasonable and efficient way of information integration. In recent years, with the continuous development of information integration technology and the formulation of a series of standards and specifications, a new solution is gradually being paid attention to, which is based on service oriented architecture the enterprise application integration (EAI) of Architecture (SOA) regards each application system in the enterprise as the service unit of SOA architecture, and establishes the enterprise application integration platform to realize the information integration between each application system. In this enterprise application integration platform, an enterprise service bus (ESB) is needed to provide standardized services. Enterprise service bus is the service operation support platform in SOA architecture, and the services encapsulated by other application systems run on this service bus, as shown in Figure 2, its establishment can effectively optimize the current enterprise ' s disordered and meshed integration mode. Secondly, we need to establish a data exchange management platform to manage all services running in the enterprise service bus and monitor the data interaction process in the integration. Finally, we need to establish a basic data management platform, as a service provider in the SOA architecture, to provide basic data management functions for other application systems. The basic data

management platform will integrate the basic data of other application systems and manage them uniformly. Other systems do not need to be managed separately.

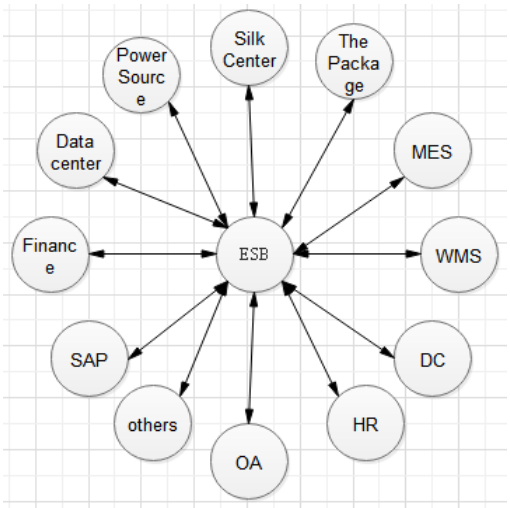


Figure 2. Schematic diagram of optimized enterprise integration channel

III. Design and Implementation

Because IBM WMB is used as the enterprise service bus, IBM DB2 will be used as the database and IBM was (WebSphere Application Server) will be used as the application server for better overall stability.

According to the demand analysis, the data exchange platform as the enterprise service bus will provide a unified entry service WS? MB. After other application systems call the service, the ESB will parse and route the called messages, and find the corresponding registered business processing WebService to call.

The data exchange management platform is responsible for the management and monitoring of IBM WMB enterprise service bus. In the data exchange management platform, it is necessary to register, modify, disable, reuse and other management functions for the service processing business. At the same time, the data exchange management platform should also realize the log recording of data sending, so as to complete the monitoring of data exchange process.

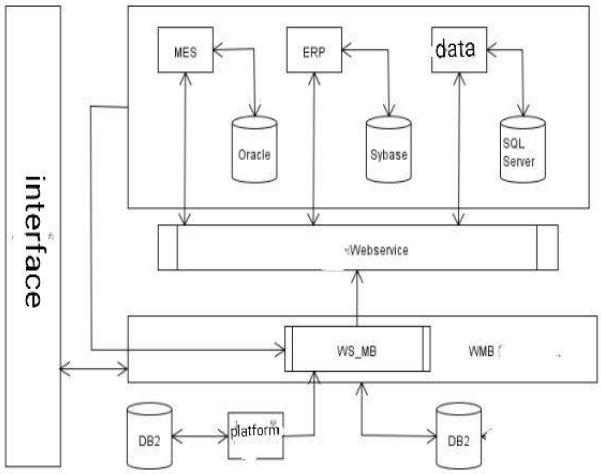


Figure 3. Technical framework of the platform

IV. Design of Enterprise Service Bus Date Exchange Platform

As the core module of enterprise application integration platform, data exchange platform needs to undertake the important work of message transmission. As shown in Figure 4, it is a basic data exchange process. The data exchange platform publishes a unified entry service WS? MB through the web service. The service caller calls the service first, and sends the call request to the data exchange platform in the form of XML message. The data exchange platform will analyze the message content, find the actual service to call, and send the message to the actual, the actual service provider will return the processed results to the data exchange platform in the form of XML message, and then the data exchange platform will return to the original caller.

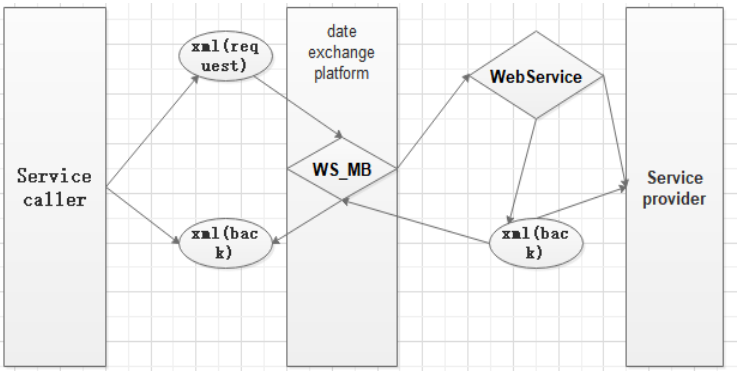


Figure 4. Flow chart of data exchange

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable the request message sent by the service caller should contain complete routing information, so how to define the routing information? In fact, from the process of data exchange, it can be seen that the three elements of data sender, data receiver and the service to be called can constitute a unique data process, so the routing information should also contain three elements.

Unified format definition of service call request XML message:

<?xml version="1.0" encoding="GB2312"?>

<Msg>

<Head>

< ID > < ID > / / message ID or serial number

< name > < name > / / message description

< source > < source > / / data source

< target > < target > / / data destination

< sername > < / sername > / / call service ID

< msgtype > < / msgtype > / / type of message (0: normal 1: request 2: answer)

< rtcode > < / rtcode > / / the return value of the corresponding request (1: success 0: failure)

< rtdesc > < / rtdesc > / / return description of the corresponding request

< backup1 > < backup1 > / / standby information

<Backup2></Backup2>

<Backup3></Backup3>

< date > XXXX / XX / XX XX: XX: XX < / date > / / message sending time

</Head>

<DATA>

<table tablename = 'table name 1' fieldname1 = 'column 1 description' fieldname2 = 'column 2 description' fieldname3 = 'column 3 description'... >

<row action = insert 'id = ' primary key value 'fieldname1 = ' AAA ' fieldname2 = ' BBB '... >

<table tablename = ' table name 2' fieldname1 = ' column 1 description' fieldname2 = ' column 2 description' fieldname3 = ' column 3 description'... >

<ROW ACTION = 'INSERT' ID = '' FIELDNAME1 = 'AAA' FIELDNAME2='BBB' .../>

</ROW>

</TABLE>

</ROW>

< / Table > / / the main body of the data sent. Table represents a data table. Row is the specific data. Table can be nested in table to represent the data of the main sub table structure.

</DATA>

</Msg>

In the XML definition, the head part describes the basic information of the data, and the three attributes of source, target and sername are the most important routing information. Through these three attributes, you can uniquely determine the service that the data needs to call, that is, the user of the service, the provider of the service and the name of the service. These attributes are registered by the service management module and are in the unified portal after receiving the XML data, the service calls the corresponding service according to the three attributes and the service registration information in the management module to send the data.

V. IMPLEMENTATION

The main functions of data exchange management platform are service management and data exchange process monitoring. As shown in Figure 5, it is the main interface of the data exchange management platform. The frequency of data exchange can be calculated through the log, and the reception and transmission volume of each system and data accessing the platform can be displayed intuitively.

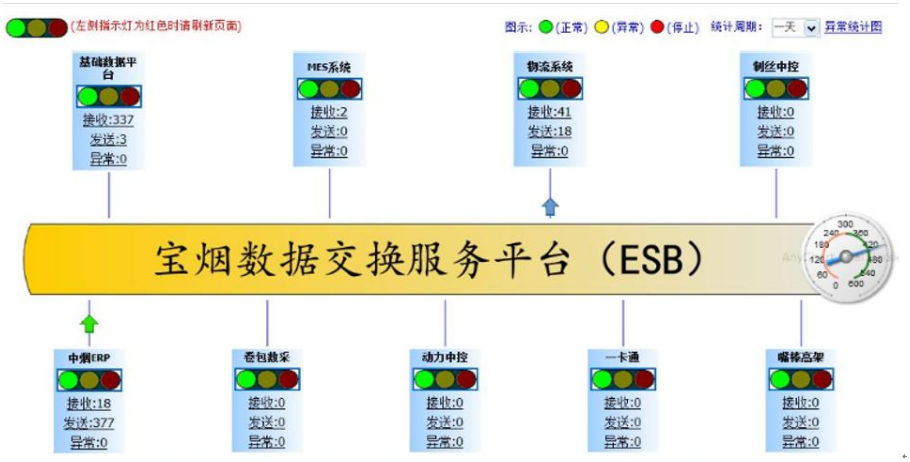


Figure 5. Main interface of data exchange management platform

a) The function of service governance is realized by registering and managing the web service published by each application system. As shown in Figure 6, the contents to be registered include: sequence number, system name, interface name, enabling tag, source, target, interface service name and WebServiceURL, namespace, calling method input object, input parameter name, output parameter name, calling method output object, extended input parameter, extended input parameter value, authentication information, WebService technology, remarks, etc. (due to confidentiality reasons, the figure is not complete).

b) Figure 7 is the implementation interface of the authority management function of the management platform for basic data. The maintenance of basic data is usually carried out by the personnel in charge of specific business. Different business personnel are usually responsible for different data. The authority management module can configure the addition, modification, deletion and query authority of various basic data according to different roles, and can also configure whether specific attributes are visible The RBAC mode is implemented.

Figure 6. Data operation module

First, configure different roles in role management, then configure the permissions of roles through role and function relationship, and finally configure the roles of different users through role and user relationship or user and role relationship. A user can have multiple roles, and a role can be played by multiple users at the same time.

Figure 8 is the implementation of the data synchronization module of the basic data management platform. By customizing the interface content and configuring different sending interfaces for different systems, you can configure whether the attribute column of the basic data is sent, the name of the sent column, etc., and you can also filter and group the sent content through SQL statements.

Figure 7. Authority management

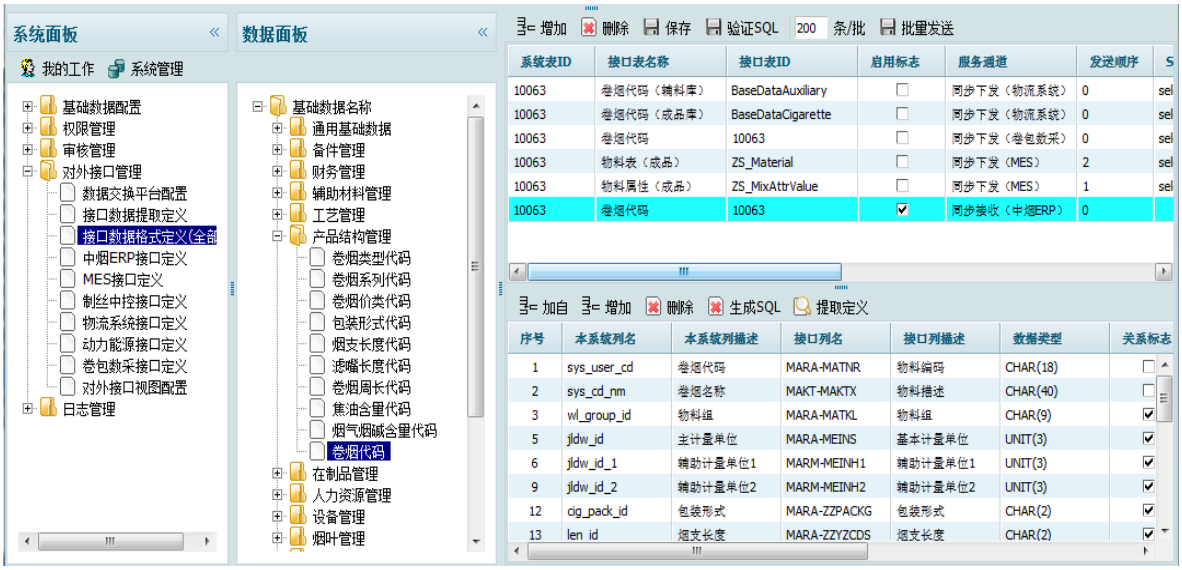


Figure 8. Definition of interface service

Other application systems publish and register the service of receiving basic data in the data exchange management platform. After the basic data management platform configures the interface, when the new basic data maintenance is completed, the platform will automatically send the data to the corresponding system through the data exchange platform according to the interface configuration. All application systems adopt this mode, and the basic data is unified.

VI. CONCLUSION

In this paper, through understanding the application status of SOA architecture in BJ cigarette enterprise, the enterprise integrated information system based on SOA architecture is proposed to solve the problems of the enterprise's current information island, high coupling between various systems, and poor integration and expansion. The basic data management platform manages and synchronizes the basic data of the whole enterprise in a centralized way, so as to solve the integration problem of application systems caused by the data inconsistency.

REFERENCES

[1] Li Zi, Yang Bin. Application of Enterprise Service Bus Technology (ESB) in large enterprises [J]. Information technology, 2013, 2:146-147.

[2] Lu Hongwei. Enterprise application integration solution based on SOA and ESB [J]. Computer application and software, 2010, 4:215-216.

[3] Chen Lingping. Design and implementation of unified application service interface platform based on SOA [J]. Network security technology and application, 2009, 4:89K. Elissa, “Title of paper if known,” unpublished.

[4] Jin Baohua, he Zhenyuan, Zhang Liang, et al. Analysis and design of data sharing and exchange platform based on SOA [J]. Journal of Zhengzhou Institute of light industry (NATURAL SCIENCE EDITION), 2011, 2:102-103.

[5] Du Wanya. Design and implementation of SOA framework based on ESB [D]. Beijing: Beijing Jiaotong University, 2008.

[6] Zhang Jun. research and implementation of distributed ESB bus based on SOA Architecture [D]. Nanjing: Nanjing University of technology, 2009.

[7] Wang Li, ye Weiquan. Planning and design of Anyan tobacco data center based on SOA Architecture [J]. Computer knowledge and technology, 2013, 1:476-478.

[8] Electronic Publication: Digital Object Identifiers (DOIs):

[9] Lin Yuqin, Huang Chenhui. Research on ESB framework for enterprise application integration [J]. Computer application, 2010, 6:1658-1660.

[10] Ma Ruijuan, he Lili. Research on data exchange service system of tobacco industry based on ESB [J]. Computer programming skills and maintenance, 2013, 6:33-35.

Comparison of Several Different Registration Algorithms

Liu Lulu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 825927856@qq.com

Liu Baolong
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: liu.bao.long@hotmail.com

***Abstract*—The common point cloud registration algorithms are usually divided into initial registration and precise registration. In this paper, SAC-IA algorithm, which is commonly used in PCL, is selected for initial registration, and the traditional ICP algorithm is used for accurate registration. Three different feature descriptors (3D shape context, Point Feature Histograms, Fast Point Feature Histograms) are used to realize SAC-IA algorithm and ICP precise registration algorithm. During the implementation of the algorithm, the registration time and registration error of point cloud are calculated; according to the experimental results, the registration time and registration error of SAC-IA algorithm and ICP algorithm based on three different descriptors are compared. The results show that the registration algorithm based on 3D shape context has high accuracy, but the registration time is too long, which is not suitable for a large number of point cloud data; the registration algorithm based on fast point feature histograms has short registration time and good registration effect.**

Keywords-Point Cloud Registration; SAC-IA Algorithm; ICP Algorithm; Registration Time; Registration Error

I. INTRODUCTION

With the rapid development of computer-aided design and computer-aided manufacturing technology, reverse engineering technology, which generates digital model through physical model, has been widely concerned. In reverse engineering, computer vision, 3D image processing and other fields, it is difficult to obtain all the data of the measured object in all directions at one time due to the influence of data acquisition equipment, the shape of the 3D object itself, external environment and so on. Usually, the point cloud data of three-dimensional objects are acquired from different angles by data acquisition equipment for many times, and the point cloud registration algorithm is used to splice the point clouds of various perspectives into the complete point cloud data. Point cloud registration is an important and difficult part of reverse engineering. The registration degree between point clouds will directly affect the accuracy of the whole 3D model, so point cloud registration has become a research hotspot in the field of point cloud processing.

Point cloud registration includes manual registration, instrument dependent registration and automatic registration, point cloud automatic registration algorithm is usually used. Automatic registration of point cloud is to calculate the dislocation between two groups of point clouds by algorithm, so as to achieve the purpose of automatic registration of point clouds. From the process of registration, it can be divided into two schemes: initial registration and accurate registration. The initial registration provides the initial transformation matrix for the accurate registration. The accurate registration is the secondary registration based on the initial transformation matrix, which can get more accurate solution and improve the final registration accuracy. Common registration algorithms include ICP algorithm[1], NDT algorithm[2], SAC-IA algorithm, etc. Among them, the accurate registration has been basically fixed to use ICP algorithm and various improved algorithms[3]~[8]. ICP algorithm has high accuracy, but it has strict requirements on the initial matrix. The results of the initial registration are not ideal, which will seriously affect the performance of the algorithm, so that the iteration cannot converge to the global optimal registration results, and even lead to the local optimal situation. Therefore, the initial registration algorithm is also very important in the registration process Important.

In this paper, SAC-IA algorithm and ICP accurate registration algorithm based on three different descriptors are selected to perform initial registration and accurate registration for two groups of point cloud

DOI: 10.21307/ijanmc-2020-004

data collected from different angles. Finally, the experimental results are compared to compare the advantages and disadvantages of several different descriptors in the initial registration algorithm and accurate registration algorithm, and the descriptor more suitable for SAC-IA algorithm and ICP algorithm are selected.

II. POINT CLOUD REGISTRATIONS

The principle of point cloud registration algorithm is to match the source point cloud Q to the reference system of the target point cloud P through the transformation matrix, that is, P = R * Q + T, where R is the rotation transformation matrix and T is the translation transformation matrix. The essence of point cloud registration algorithm is the process of solving R and T. The specific implementation steps are as follows:

Step1. Extract key points from two sets of point cloud data sets according to the same key point selection criteria;

Step2. Calculate the feature descriptors of all the selected key points;

Step3. Combined with the coordinate position of the feature descriptors in the two sets of point cloud data sets, based on the similarity of the features and positions between the two sets of point cloud data sets, the corresponding relationship between them is estimated, and the corresponding point pairs are preliminarily estimated;

Step4. For the noise problem of point cloud data, remove the wrong corresponding point pairs that have influence on registration;

Step5. Use the residual correct correspondence to estimate the rigid body transformation and complete the registration.

III. SAMPLE CONSENSUS INITIAL ALIGNMENT

The initial registration is to prepare for the subsequent accurate registration. The initial registration is carried out for two pieces of point clouds, and the initial values of translation matrix and rotation matrix are calculated. Then the point cloud data to be registered is transformed into a unified coordinate system, providing a better initial position for accurate registration.

For the rough estimation of the initial transformation matrix, greedy initial registration method has a lot of work, using the point cloud data rotation invariant feature, and the computational complexity is high, so it is necessary to check all possible correspondence of the feature descriptors; in addition, greedy algorithm may fall into the local optimal solution. Therefore, for the initial registration method of point cloud, we usually choose the sampling consistency method to try to maintain the geometric relationship of the same correspondence, rather than trying to understand all combinations of finite correspondence. Sample Consistence Initial Alignment (SAC-IA for short) algorithm takes a large number of samples from the candidate correspondence, and quickly finds a good transformation by looking at a large number of correspondences. Until the best rotation and translation errors are obtained and stored.

In the initial registration algorithm of point cloud, 3D point cloud feature description and extraction is the most basic step, and also the most critical part of the initial registration algorithm of sampling consistency. Sampling consistency registration algorithm is based on local feature description. This chapter mainly realizes the initial registration algorithm of sampling consistency based on three descriptors: 3D shape content descriptors, point feature histogram descriptors and fast point feature histogram descriptors, the optimal results of the initial registration algorithm are obtained by experiments.

A. 3D shape context

3D shape context (3dsc for short) uses a vector to describe the shape features of the specified points and their fields on the surface, and establishes the corresponding relationship between the points of different surfaces by matching the values of the vector, which is the descriptor of the specified points. 3Dsc descriptors are simple in structure, strong in discrimination and insensitive to noise. The construction method is as follows: in the spherical support domain with designated point P as the center, the grid is divided into three coordinate directions: radial direction, direction angle and pitch angle, the number of points falling into the grid is counted, and the vector V is constructed. Each element of V corresponds to a grid in the support domain. The value of the element is the number of points in the corresponding grid. The vector V is the descriptor of point P[10]. 3Dsc grid division is shown in Fig. 1.

B. Point Feature Histograms

Point Feature Histograms (PFH for short) refer to the spatial differences between query points and neighboring points by parameterization, and form a multi-dimensional histogram to describe the neighbor geometric properties of point k. The high-dimensional hyperspace of histogram provides a measurable

information space for feature representation, which is invariant to the 6-dimensional pose of the corresponding surface of point cloud, and robust under different sampling density or neighborhood noise level. PFH representation is based on the relationship between points and their k neighbors and their estimated normal, considering all the interactions between normal directions, trying to capture the best change of sample surface to describe the geometric characteristics of samples. Therefore, the synthetic feature hyperspace depends on the quality of the surface discovery estimation of each point. Fig. 2 shows the influence area of PFH calculation of a query point (P_q). P_q is marked in red and placed in the middle of the ball. The radius is r, and all K-neighbor elements of P_q (all points whose distance from point P_q is less than radius r) are all connected in a network. The final PFH descriptor get histogram by calculating the relationship between all two points in the domain.

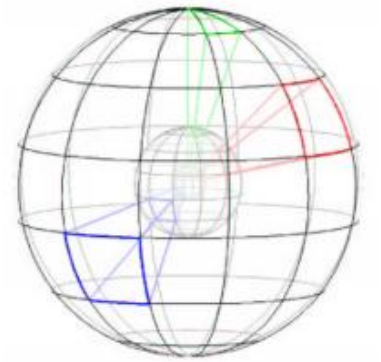


Figure 1. 3Dsc mesh generation

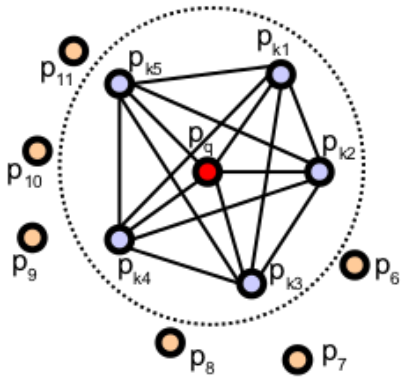


Figure 2. Influence range chart of PFH calculation at query point P_q

C. Fast Point Feature Histograms

Fast point feature histograms[11](FPFH for short) is a feature descriptor based on the normal angle between points and their neighbors and the angle between the lines between points. It is an improved algorithm by PFH, which retains the main geometric characteristics of point-to-point description in PFH, and reduces the computational complexity from $O(nk^2)$ to $O(nk)$, where n is the number of points in the point cloud data, and k is the number of neighbors considered when calculating the eigenvector of each point.

For a persistent query point P_q, FPFH first uses the pairing between P_q and its neighborhood points (represented by the red line in Fig. 3) to estimate its Simplified Point Feature Histograms (SPFH) value. Compared with the standard calculation of PFH, FPFH has less inter neighborhood interconnection. All points in the point cloud data set need to calculate and obtain SPFH, and calculate the weight according to the SPFH value of the adjacent point of P_q point and the SPFH value of P_q point, and finally get the FPFH value of P_q point. For the calculation connection pairs added in FPFH calculation, the black line in Fig. 3 indicates that some important point pairs (points directly connected with P_q points) are represented by repeated cardinality twice (the thick line in Fig. 3 indicates), and other connected points are represented by thin black line.

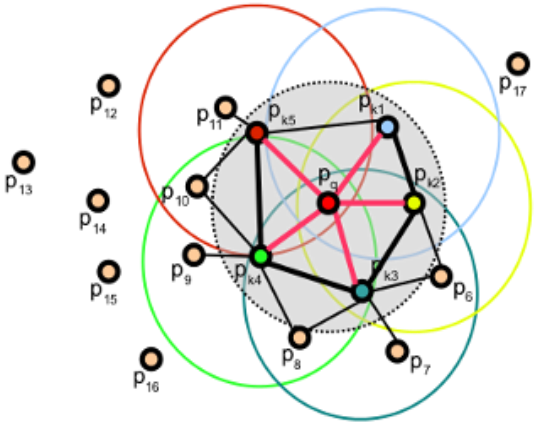
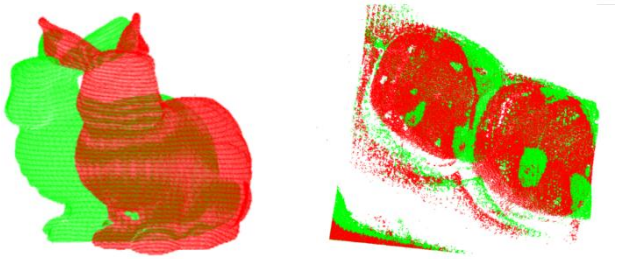


Figure 3. FPFH calculation neighborhood influence range of query point P_q

D. Experimental Verification

In this experiment, the classic Bunny point cloud model of Stanford University (as shown in Fig. 4 (a)) is used to compare the efficiency and accuracy of the three algorithms. Finally, the tooth point cloud collected by the laboratory point cloud acquisition equipment (as shown in Fig. 4 (b)) is registered to verify the feasibility of the algorithm comparison results. In Fig. 4, the green point cloud is the source point cloud and the red point cloud is the target point cloud. In Bunny point cloud model, the number of source point cloud is 40256, and the number of target point cloud is 35336; the number of tooth point cloud

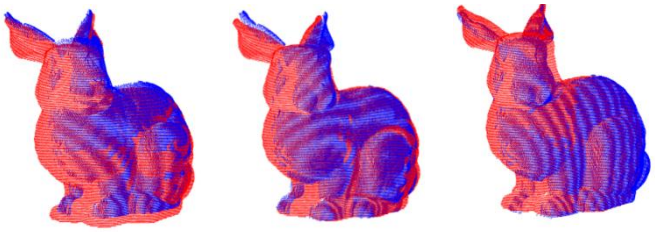
data source point cloud collected in laboratory is 140450, and the number of target point cloud is 145366.



(a) Bunny model (b) Tooth point cloud model

Figure 4. Point cloud data model

The general evaluation standard of point cloud registration is LCP (Largest Common Pointset). That is to say, given two groups of point set P and Q, a transformation T (P) is found, which makes the maximum overlap of P and Q after transformation. If there is another Q point in the tolerance range at any point in the transformed P, it is considered to be a coincidence point. The proportion of coincident points to the number of all points is the degree of overlap. In this paper, the registration accuracy is determined by the rotation and translation errors and distance errors of the registration point cloud relative to the target point cloud on the x-axis, y-axis and z-axis. Compared with LCP, the registration accuracy of the point cloud can be more intuitively expressed. The experimental results of SAC-IA algorithm registration based on 3Dsc, PFH and FPFH feature descriptors are shown in Fig. 5, among them, the red point cloud is the target point cloud, and the blue point cloud is the registered point cloud.



(a) 3Dsc+SAC-IA (b) PFH+SAC-IA (c) FPFH+SAC-IA

Figure 5. Initial registration results

The registration time, rotation angle error on XYZ axis, translation distance error and distance error between registration point cloud and target point cloud of two sets of bunny point cloud are shown in Table I.

TABLE I. INITIAL REGISTRATION ERROR TABLE

Comparison of registration algorithms	**Feature Descriptors**		
	3Dsc	***PFH***	***FPFH***
SAC-IA runtime	417.381 s	22.727 s	8.085 s
x-axis rotation error	0.0400952 °	0.0474209 °	0.0265896 °
y-axis rotation error	0.811471 °	0.758933 °	0.709945 °
z-axis rotation error	-0.740596 °	-0.734409 °	-0.756489 °
x-axis translation error	0.012691 mm	0.007121 mm	0.019262 mm
y-axis translation error	-0.297463 mm	-0.29991 mm	-0.298785 mm
z-axis translation error	-0.192225 mm	-0.194115mm	-0.194264 mm
distance error	0.354395 mm	0.357320 mm	0.356906 mm

In terms of algorithm efficiency, 3Dsc needs to calculate the surface shape characteristics of point cloud, which increases the calculation amount; compared with PFH and FPFH, it takes a long time, and FPFH is an improved algorithm based on PFH, that is, FPFH algorithm is more efficient and takes the shortest time. In terms of algorithm accuracy, the smaller the rotation, translation and distance errors of x-axis, y-axis and z-axis are, the higher the accuracy of the algorithm is. From Table I, it can be seen that the accuracy of 3Dsc algorithm combined with SAC-IA algorithm is higher than that of FPFH and PFH algorithm. Therefore, the experimental results show that the algorithm based on FPPH feature descriptor has the shortest registration time and the highest efficiency, and the algorithm based on 3Dsc feature descriptor has relatively high accuracy.

IV. ITERATIVE CLOSEST POINT

Through the initial registration, the two sets of point cloud data roughly coincide, but the registration accuracy is still far from the requirements of practical applications. Therefore, accurate registration of point cloud data is required to reduce registration errors. In order to register the two groups of point cloud as much as possible and minimize the error between them, this paper uses the classic Iterative Closest Point (ICP for short) algorithm for accurate registration.

ICP algorithm[12][13] is the mainstream algorithm for 3D model registration. For each point in the source point cloud, an exhaustive search method is used in the target point cloud to search for the closest point as the

corresponding point; then the transformation matrix of all corresponding point pairs is registered and aligned, and the source point cloud is finally calculated according to The obtained transformation matrix is transformed. If the measurement error is not considered, the accuracy of the ICP algorithm is affected by the measurement sampling density, and the error value is proportional to the average sampling distance. That is, the higher the sampling density, the higher the accuracy of the stitching.

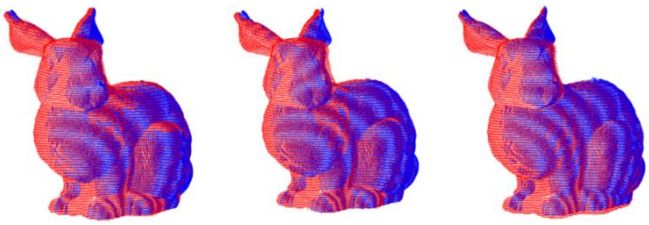
The basic principle of ICP algorithm is to find the nearest point (P_i, Q_i) in the target point cloud P and source point cloud Q to be registered according to certain constraints, and then calculate the optimal matching parameter rotation matrix R and translation matrix T to minimize the error function. The error function $F(R,T)$ conforms to equation (1).

$$F(R,T)=\frac{1}{n}\sum_{i=1}^{n}\| Q_i-(RP_i+T)\|^2 \qquad (1)$$

Where n is the number of nearest neighbor point pairs, P_i is a point in the target point cloud P, $Q_i$is the nearest point corresponding to P_i in the source point cloud Q, R is the rotation matrix, and T is the translation matrix.

A. Experimental Verification

According to the initial registration results, the point cloud data is accurately registered by the ICP algorithm. The SCA-IA algorithm based on the 3Dsc, PFH, and FPHF feature descriptors combined with the accurate results of ICP is shown in Fig. 6.



(a) 3Dsc+ICP (b) PFH+ICP (c) FPFH+ICP

Figure 6. Accurate registration results

After ICP precise registration, the initial registration time, precise registration time, rotation angle error on XYZ axis, translation distance error, and the distance error between the two point clouds of the registration point cloud and the target point cloud are as follows: Table II.

TABLE II. ACCURATE REGISTRATION ERROR TABLE

Comparison of registration algorithms	**Feature Descriptors**		
	3Dsc	***PFH***	***FPFH***
total time	415.765 s	23.804 s	9.094 s
SAC-IA runtime	414.86 s	23.15 s	8.435 s
ICP runtime	0.905 s	0.654 s	0.659 s
x-axis rotation error	0.0141387 °	0.0325149 °	0.0134278 °
y-axis rotation error	0.764633 °	0.721214 °	0.744733 °
z-axis rotation error	-0.758194 °	-0.76543 °	-0.770207 °
x-axis translation error	0.0128116 mm	0.013648 mm	0.015574 mm
y-axis translation error	-0.298834 mm	-0.29914 mm	-0.299785 mm
z-axis translation error	-0.194931 mm	-0.19437 mm	-0.193715 mm
distance error	0.357021 mm	0.357004 mm	0.357266 mm

As can be seen from Table II, in terms of algorithm efficiency, the SAC-IA algorithm based on the FPFH feature descriptor combined with the ICP algorithm is relatively shorter. From the perspective of algorithm accuracy, the registration accuracy of SAC-IA algorithm based on 3DSC, PFH, FPFH feature descriptors and ICP precise registration algorithm is not much different. Therefore, for the dental point cloud model collected in the laboratory, the accuracy of the three descriptors is not much different, but in terms of efficiency, the FPFH descriptor is more suitable for point cloud data with a large amount of data. Therefore, the tooth model is registered using the SAC-IA algorithm based on the FPFH feature descriptor and the ICP precise registration algorithm. The registration result is shown in Fig. 7.

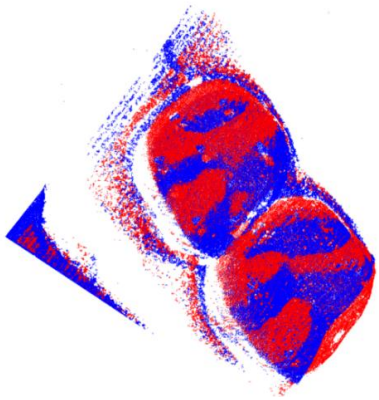


Figure 7. Tooth Model Registration Results

V. CONCLUSION

In this paper, based on the SAC-IA algorithm of 3Dsc, PFH and FPFH, the bunny point cloud is registered, and the efficiency and accuracy of the initial registration algorithm are analyzed and compared. In the initial registration, SAC-IA algorithm based on 3dsc feature descriptor has a relatively high accuracy, but it takes a long time, which is suitable for the registration with a small number of point clouds and high accuracy requirements; SAC-IA algorithm based on FPFH feature descriptor has a high efficiency, which is suitable for the registration with a large number of point clouds and high efficiency requirements. For the accurate registration algorithm, the ICP algorithm of PCL is used in this paper, and the effect is not ideal from the registration results; in addition, the ICP algorithm can be improved to improve the overall accuracy of the registration algorithm.

ACKNOWLEDGMENT

The successful completion of this paper cannot be separated from the help of teachers, students and funds. Thanks to Professor Liu Baolong, Mr. Wu Qiong and Mr. Si Lipeng for their guidance and help. Finally, I would like to thank the science and technology program of Weiyang District of Xi'an City (Project No.: 2018036) for its fund support.

REFERENCES

[1] Besl P J, Mckay H D. A method for registration of 3-D shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2):0-256.

[2] Peter Biber and Wolfgang Straßer. The normal distributions transform: A new approach to laser scan matching. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), pages 2743–2748, Las Vegas, USA, October 2003.

[3] Greenspan M, Yurick M. [IEEE Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. - Banff, Alberta, Canada (6-10 Oct. 2003)] Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings. - Approximate K-D tree search for efficient ICP[J]. 2003:442-448.

[4] Andreas Nüchter, Lingemann K, Hertzberg J . Cached k-d tree search for ICP algorithms[C]// Sixth International Conference on 3-D Digital Imaging and Modeling, 3DIM 2007, 21-23 August 2007, Montreal, Quebec, Canada. IEEE, 2007.

[5] Timothée Jost, Heinz Hügli. Fast ICP Algorithms for Shape Registration[M]// Pattern Recognition. Springer Berlin Heidelberg, 2008.

[6] Silva L, Bellon O R, Boyer K L. Precision range image registration using a robust surface interpenetration measure and enhanced genetic algorithms. [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2005, 27(5):762-776.

[7] Du S, Liu J, Zhang C, Zhu J, & Li K. Probability iterative closest point algorithm for m-D point set registration with noise[J]. Neurocomputing, 2015, 157:187-198.

[8] Chetverikov D, Stepanov D, Krsek P. Robust Euclidean alignment of 3D point sets: the trimmed iterative closest point algorithm[J]. Image and Vision Computing, 2005, 23(3):299-309.

[9] Zhu Dehai. PCL course of dianyun Library[M]: Beijing University of Aeronautics and Astronautics, 2012.

[10] Frome A, Huber R, Kolluri T, Buelow M.J. Recognizing Objects in Range Data Using Regional Point Descriptors. Proceedings of the European Conference on Computer Vision (ECCV), Prague, Czech Republic, 11–14 May 2004; 3, pp. 224–237.

[11] Rusu R B, Blodow N, Beetz M. Fast Point Feature Histograms (FPFH) for 3D registration[C]// Robotics and Automation, 2009. ICRA '09. IEEE International Conference on. IEEE, 2009.

[12] Zhang Z. Iterative point matching for registration of free-form curves and surfaces[J]. International Journal of Computer Vision, 1994, 13(2):119-152.

[13] Chen Y, Gérard Medioni. Object modeling by registration of multiple range images[J]. Image and Vision Computing, 1992, 10(3):145-155.

Comparison Research on Future Network Between IPv4, IPv6 and IPV9

YURY Halavachou
Department of the International Relations
Belarusian State University of Transport
Republic of Belarus
34, Kirova street, Gomel, 246653
Republic of Belarus
E-mail: oms@bsut.by

Xu Fei
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: xinfei2000@qq.com

***Abstract*—IPv4 is the most widely used protocol on the Internet, and its address space is 232. In the early stage of the Internet, due to the underestimation of the development of Internet, IP resources were very limited. By 2010, there was no address that could be allocated. In order to solve the problem of insufficient addresses, the IETF designed the next generation IPv6 protocol to replace IPv4, IPv6 has 2128 addresses in theory, and however, only one-eighth of the addresses can actually be allocated to end users. At present, 128 barcodes are already having 128 bits, and it cannot be covered, so IPv6 have some considerable limitations. In 1998, Chinese researcher proposed IPV9. In order to distinguish from IPv4 and IPv6, the V in IPV9 is uppercase, not lowercase. The IPV9 includes three technologies: a new address coding design, a new addressing mechanism and a new address architecture design. These technologies constitute the core technology system at the bottom of the new generation IP network. The new network framework designed on this basis can form a network system that is connected and compatible with existing networks. IPV9 is not a simple upgrade of IPv4 or IPv6 and its address space is 10256 by default. The massive address can meet the needs of human activities about 750 years, this paper will introduce the characteristics of the future network, and comparison the system between the IPv4, IPv6 and IPV9, and lay a solid foundation for the subsequent development.**

Keywords-IPv4; IPv6; IPV9; Future Network

With the rapid development of science and technology, the world has entered an information age of data communication. The most famous of the data networks is Internet, a packet-switched network developed by the U.S. department of defense called the ARPANET, which is considered the precursor to the information superhighway. Now almost all countries and regions have joined the Internet.

In order for information to be transmitted correctly over the Internet to its destination, each computer connected to the Internet must have a unique address. At present, there are three kinds of address compilation methods: one is "IP address", which consists of four digits divided by the dot; the other is "domain name", a series of strings split by dots, and the third is "Chinese domain name system", which consists of a three-level domain name split by a decimal and an oblique line. These three address structures have become the current network system; bringing great convenience to people's

DOI: 10.21307/ijanmc-2020-005

access to the Internet, the network has completely changed people's lives.

I. PROBLEMS WITH IPV4 AND IPV6

A. Problems with IPv4

1) Insufficient address

Internet uses IPv4 protocol address scheme, the address number up to 232, due to the early development of the Internet to estimate the development trend of Internet, the IP allocation is not reasonable, address resources are exhausted, although no classification of addressing CIDR technology, network address translation NAT technology to alleviate the crisis, but still can't solve the problem. And address will be more and more widely used in e-commerce logistics code, space code, identity code, digital currency, three-dimensional geographical code and other intelligent terminals, the original address allocation technology cannot meet the needs of social development.

2) Route table expansion

The topology of address space directly results in the form of address allocation independent of the network topology. With the increase of the number of networks and routers, the excessively expanded routing table increases the search and storage overhead and becomes the bottleneck of the Internet. At the same time, the length of packet head is not fixed, and it is very inconvenient to extract, analyze and select routes by hardware, so it is difficult to improve the throughput of routing data. Then there is the uneven distribution of IP addresses. Due to its origin in the United States, more than half of all addresses are owned by the United States, resulting in a serious imbalance in the distribution of IP addresses.

3) Lack of quality of service (QoS)

IPv4 was originally designed for military use, and was not intended to be open to the outside world. As a result, QoS of quality of service and security were very poor, and it was difficult to provide rich QoS functions for real-time multimedia, mobile IP and other commercial services. Although the later developed protocols such as RSVP provided QoS support, the cost of planning and constructing IP network was relatively high.

Despite of its shortcomings, IPv4 was the first network all over the world, and people had got to used it, so it will going forever.

B. Problems with IPv6

The length of IPv6 is 128 bits, or 2128 addresses. The address space is much larger than the 32-bit address space. Moreover, the principle of Aggregation is adopted, which enables the router to represent a subnet with an Entry in the routing table, it greatly reducing the length of routing table in the router and improving the speed of forwarding packets. The addition of Multicast support and Flow Control over IPv6 has led to significant advances in multimedia applications, providing a good network platform for, Quality of Service Control (QoS). Despite its obvious advantages, IPv6 has a big flaw in the design of its address structure. The shortcomings are as follows.

1) Structural hierarchy disorder

IPv6 confuses the network hierarchy in the design, and the interface ID inserts the physical address into the logical address layer, which on the one hand results in the physical address space forming a limitation on the empty IP address, the security does not belong to the content of the IP layer, it is not necessary to design security technology in the IP layer. Because with the development of security technology, security methods and key length will change constantly, so the development of security technology will eventually lead to the need for IP address redesign.

2) Ambiguous address space

In the unicast address with more IPv6 applications, the structure of "network ID+ host ID" similar to IPv4 is adopted from a large point of view, and the network

ID of IPv6 is changed into a three-layer more structure with a fixed length of subnet prefix: "top-level aggregation ID+ secondary aggregation ID+ site-level aggregation ID". IPv6 is a kind of patchwork addressing. So its address space is not pure 128 bits.

IPv6 address space is not the 128-bit address space that people think of. Due to the special address structure design, IPv6 itself has to go through three significantly different version transitions if it wants to truly implement the 128-bit address space, the IPv16 for 64-bit effective address space; IPv26 for 128-bit valid address space. The transition between the three versions is like to upgrade the three different protocols.

3) Incompatible with IPv4

IP address is the basic protocol of the Internet, and it is very difficult to solve it through complete replacement. Initially, without further study, the designers of IPv6 decided that the 32-address space problem of IPv4 could not be solved by a smooth upgrade, so they simply redesigned it entirely from scratch. IPv6 requires all nodes of the entire network to support the new IP protocol, and all terminal operating systems and applications to support upgrades, making the problem extremely difficult.

These shortcomings are also the main reason why IPv6 has not been widely used since its emergence.

II. FUTURE NETWORK IPV9

A. Process of IPV9

In December 1998, Xie Jianping, a scholar from Shanghai, China and the inventor of the Future Network, applied to the National Intellectual Property Administration (NIPA), PRC (formerly the Patent Office of China) for the invention patent of "the method of assigning addresses to computers connected to the network with full digital codes", which was officially authorized by the NIPA on November 7, 2001.

In December 1998, Mr. Xie jianping registered the copyright in the national copyright administration of China in “the method of unified compilation and distribution of addresses of networked computers and intelligent terminals”, “the overall distribution method of computer addresses allocated by full decimal algorithm for networked computers”, and “the gateway of decimal number”.

In October 2001, the “copyright of IPV9 protocol and application” was registered.

In 2001, the former Science and Technology Department of the Ministry of Industry and Information Technology of China established the China Decimal Network Standards Working Group (IPV9 Working Group) with enterprises as the main body and industry, university and research institute as a combination.

In 2002, the “code for digital domain names” was published, defining the "decimal network, IPV9 resource record and management organization".

In 2007, the former Ministry of Industry and Information Technology of China formally defined IPV9 as the "future network" to distinguish the next generation of the Internet for IPv6.

In 2011, the authoritative professional institutions of the US government have confirmed legally and technically that China has the core technology of sovereign network with independent intellectual property rights under the IP framework. This is the patented technology of IPV9 which is different from the existing technology of the US Internet. The official patent name is "Method of using whole digital code to assign address for computer".

In December 2011, the U.S. federal patent and trademark office issued a patent certificate numbered US 8,082,365, stating in its notice of approval that the applicant's identification report was "very convincing".

On May 21, 2013 and March 11, 2014, the United States twice voted in favor of the China-led "naming and addressing" and "security" of the future network.

On February 23, 2013, the State Council issued the national science and technology infrastructure construction medium and long term plan (2012-2030), in order to break through the future network basic theory and support the new generation of Internet experiments, the construction of future network test facilities.

On June 1, 2016, the Ministry of Industry and Information Technology of China released relevant industry standards for IPV9 implemented nationwide: Including SJ/T11605 "for products and services based on the technology of radio frequency domain rules", "SJ/T11604 decimal network based RFID tag information orientation, query and service discovery technology standard", SJ/T11603 "used Digital ID format in information processing products and services", SJ/T11606 "the network architecture of RFID tags information query service specification", SJ/T11682 "based on the electronic tag information of decimal network location, query and service discovery and application".

B. About IPV9

IPV9 is completely independent intellectual property rights on the basis of full decimal digit code, it has 2^{256} of cyberspace sovereignty, including from mother root, master root, 13 root name servers, using zero trust security communication mechanism after verification first, compatible with the current Internet system, with overlapping geographical position and the IP address space for the future network architecture.

On the basis of compatibility with all the functions of the Internet at present, IPV9 adopts the TCP/IP/M three-layer and four-layer hybrid architecture, with mixed virtual and real circuits, to complete the video data transmission of large code stream.

IPV9 obtained Chinese patent in 2001 (CN98 1 22785), and has obtained authorized patents successively in more than ten countries and regions, including South Africa, Turkey, Kazakhstan, Russia, South Korea, North Korea, Hong Kong, Canada, Singapore, Australia, Mexico and Norway. IPV9 applied for US patent in 2004. It was issued seven times of "non-final rejection opinion" and six final rejections by the US Patent Office. During this period, it was repeatedly criticized by senior members of the US IETF and famous American IT companies. In December 2011, the US Patent and Trademark Office officially issued a patent certificate numbered US 8,082,365, and clearly stated in its approval notice that the appraisal report provided by the applicant was “very convincing”. In December 2011, the US Patent and Trademark Office officially issued a patent certificate numbered US 8,082,365, and clearly stated in its approval notice that the appraisal report provided by the applicant was "very convincing".

III. SPECIAL CHARACTERISTICS OF IPV9

1) Address space is huge

IPV9 has a larger address space than IPv4/IPv6. IPv4 defines the bit length of IP address is 32, that is, there are 232-1 addresses; While the length of IPv6 is 128, that is, 2128-1 addresses, the standard length of an IPV9 address is 2256-1, with 42 layers address structure design will be 10256-1 (21024-1). To put it mildly, if IPv6 were widely used, every grain of sand in the world would have an IP address. Then after IPV9 is widely used, the smallest molecule of bright matter in the whole universe will have a corresponding address. It is no exaggeration to say that if IPV9 is fully applied, every cell and living gene in the world can be assigned to an IPV9 address. Layer 42 is the asset management address (including legal digital currency space) compatible with ean-ucc128 barcode length.

2) Route tables are smaller

IPv6 has a smaller routing table than IPv4. The address allocation of IPv6 follows the principle of Aggregation at the beginning, which enables the router to represent a subnet with an Entry in the table, this greatly reducing the length of routing table in the router, and improving the speed of forwarding packets in the routing table.

The routing table of IPV9 is very small, and the address allocation of IPV9 follows the principle of Geo-spatial clustering from the beginning, which enables IPV9 router to represent a country subnet and an application subnet with a single record, it greatly reducing the length and cleanliness of routing table in the router, and improving the speed of forwarding packets by routing table. At the same time, this subnet can express a specific geographical location, for example, we assign the IPV9 address segment of Shanghai as 86[21[5]/96, then in other routers of the same level, only one route pointing to the address segment of 86[21[5]/96 can realize the IPv9 address routing of Shanghai. According to this logic, only one route is needed from country to country. For example, the route to China is 86/64. The IPv4 routing table is large and irregular, and the IPv6 routing table is smaller than IPv4, but the IPv6 routing table contains no geographic information and the routing is messy.

3) Automatic configuration support

IPV9 adds support for automatic configuration of variable length addresses, which is an improvement and extension of DHCP protocol of IPV9, making network management more convenient. IPV9 supports multicast, and supports the ISO/IEC C6 future network << naming and addressing >>TCP/IP/M model, and supports long packet code streams for virtual and real circuits. This allows multimedia applications on the web to ensure video quality and reduce overhead, provide faster and faster applications such as industrial controls and unmanned vehicles, and provide better and cheaper service over the Internet than IPv6.

4) Address length could be select

IPV9 address length has a variety of options, which can realize the change of 16, 32, 64, 128, 256, 512 and 1024 bit address length, and select the most appropriate address length according to different usage scenarios to reduce the routing overhead.

5) Dual encryption

The address length of IPv9 is long enough to realize dual encryption from the transmission of source and target addresses, which plays an important role in some specific network transmission fields.

6) Add location information to the address

IPV9 addresses can be embedded with geo-location information, as well as personal and industry ID information, this making IP addresses uniquely tied to personal information.

7) Compatible with previous addresses

IPV9 address is backward compatible with IPv4/IPv6 address. In order to absorb the upgrade difficulty of IPv6 incompatibility with IPv4, IPV9 protocol remains and unchanged, so that IPv4/IPv6 upgrade to the new version of IPV9, the upgrade cost is very low.

8) Sovereignty is different

IPv4/IPv6 addresses Spaces and copyright ownership: United States.

IPV9 address space and copyright ownership: China.

IV. FEATURE OF IPV9

IPV9 technology has many features; a comparison of IPV9 and IPv4, IPv6 features is listed below.

TABLE I. COMPARISON BETWEEN IPV4 AND IPV9

Item	IPv4	IPV9
Bit length	32	256
Address format	Dot decimal, uncompressible	[ ] Bracket decimal notation, with zero compression, can be compressed on both sides
Network express	Mask or length prefix representation	Length prefix express that supports public geographic space clustering
Loop Address	127.0.0.1	[7]1
Public address	Common public IP address	Aggregate global address location unicast addresses
Automatic configuration	Automatically configured address（169.254.0.0/16）	Link-Local Address:4269801472[0/64
Broadcast address	Contains broadcast address	No broadcast address, transitional support broadcast address
Unspecified address	0.0.0.0	[8]
Domain name resolution	IPv4 Host address(A) resource record	IPv9 host address (AAAAAAAA) resource record
Mother root server space	32bits (232-1 addresses)	Realized 256bits (2256-1 addresses) design objective 2048bits
Root domain server name	13 letters from A to M	13 letters from N to Z
China top-level domain	.CN	.CHN
Inverse Resolution	IN-ADDR.APRA Domain	IN-ADDR.APRA9 Domain
Compatibility 1	Incompatible with IPv6 addresses	Compatible IPv6 address：y]y]y]x:x:x:x:x:x:d.d.d.d
Compatibility 2	Incompatible with IPV9 addresses	Compatible IPv4 address：y]y]y]y]y]y]d.d.d.d
Transition address	No	Transition address IPv4：[7]d.d.d.d 简写 J.J.J.J
Encryption	No IP address encryption	Dual encrypted of the source address and the destination address
Address length	Fixed 32 bits	Not fixed，canbe16、32、64、128、256、512、1024bits
Geographic information	No geographic location information	Geographic location information Can be embedded
DHCP	Nonsupport DHCP	Added support for automatic configuration of variable - length addresses
ISO/IEC C6 & TCP/IP/M model	Not supported	Supported
Communication rules	Communicate first, then verify	Verify before communication
Network model	TCP/IP	TCP/IP/M
Sovereign	America	China

TABLE II. COMPARISONS BETWEEN IPV6 AND IPV9

Item	IPv6	IPV9
Bit length	128	256
Address format	Colon-separated hexadecimal with zero compression, single compression	[ ] [ ] Bracket decimal notation, with zero compression, can be compressed on both sides
Network express	Mask or length prefix representation	Length prefix express that supports public geographic space clustering
Loop Address	：：1	[7]1
Public address	Can aggregate the global single point transmission address	Aggregate global address location unicast addresses
Link-Local Address	FE80：：/64	4269801472[0/64
Broadcast address	No	No broadcast address, transitional support broadcast address
Unspecified address	0: 0: 0: 0: 0: 0: 0: 0	[8]
Domain name resolution	IPv6 Host address(AAAA) resource record	IPv9 host address (AAAAAAAA) resource record
Mother root server space	128bits (2128-1 addresses)	Realized 256bits (2256-1 addresses) design objective 2048 bits
Root domain server name	13 letters from A to M	13 letters from N to Z
China top-level domain	.CN	.CHN
Inverse Resolution	IP6.INT Domain	IN-ADDR.APRA9 Domain
Compatibility 1	Incompatible with IPv9 addresses	Compatible IPv6 address：y]y]y]x:x:x:x:x:x:d.d.d.d
Compatibility 2	Incompatible with IPv4 addresses	Compatible IPv4 address：y]y]y]y]y]y]d.d.d.d
Transition address	No	Transition address IPv4：[7]d.d.d.d 简写 J.J.J.J
Encryption	No IP address encryption	Dual encrypted of the source address and the destination address
Address length	Fixed 128 bits	Not fixed，canbe16、32、64、128、256、512、1024bits
Geographic information	No geographic location information	Geographic location information Can be embedded
DHCP	Support DHCP, no automatic configuration for variable-length addresses	Added support for automatic configuration of variable - length addresses
Network model	TCP/IP	TCP/IP/M
ISO/IEC C6 & TCP/IP/M mode	Not supported	Supported
Communication rule	Communicate first, then verify	Verify before communication
Network model	TCP/IP	TCP/IP/M
Sovereign	America	China

V. CONCLUSIONS

The IPV9 protocol uses 0-9 Arabic digital network as the virtual IP address and uses decimal as the text representation method, which is a convenient way to find online users. IPV9 has a large number of assignable IP addresses, and the maximum number of address bits is 2×1048 In order to improve efficiency and facilitate end users, some addresses can be used directly as domain names, which is the cornerstone of the future digital world. At the same time, IPV9 is also called "New Generation Security and Reliable Information Integrated Network Protocol" because it uses the classification and coding of the original computer network, cable broadcast television network and telecommunication network.

IPV9 technology and the whole network architecture make China to be the second country in the world with complete future network architecture. This paper introduces the generation process and characteristics of IPV9, and compared with the existing Internet, with the continuous optimization and improvement of IPV9 Future Network, it will be applied in many other countries.

REFERENCES

[1] Xie Jianping etc.Method of using whole digital code to assign address for computer [P]. US: 8082365, 2011.12.

[2] S. Deering, R. Hinden, Network Working Group. Internet Protocol, Version 6 (IPv6)-Specification, RFC-1883, 1995.12.

[3] M. Crawford. Network Working Group. Transmission of IPv6 Packets over Ethernet Networks.RFC-2464, 1998.12.

[4] J. Onions, Network Working Group. A Historical Perspective on the usage of IP version 9. RFC1606. 1994.04.

[5] Cerf, Network Working Group. A VIEW FROM THE 21ST CENTURY, RFC1607. 1994.04.

[6] Information technology-Future Network- Problem statement and requirement-Part 2: Naming and addressing, ISO/IEC DTR 29181-2, 2014, 12.

[7] Radio frequency identification tag information query service network architecture technical specification. SJ/T11606-2016, 2016. 06

[8] Soliman H., Castelluccia C., Malki K. E., Bellier L. Hierarchical Mobile IPv6 (HMIPv6)

[9] Johnson D., Perkins C., Arkko J. Mobility Support in IPv6," RFC 3775, June 2004.

[10] Perkins C.E. IP Mobility Support for IPv4", IETF RFC 3220, Jan. 2002

[11] Farinacci D., Fuller V., Meyer D., Lewis D. Interworking LISP with IPv4 and IPv6", draft-ietf- lisp-interworking-01, Aug. 2010

[12] Meyer K.F.D., & Zhang L. RFC 4984: Report from the IAB Workshop on Routing and Addressing, September 2007.

[13] Dommety G., & Jain R. Potential networking applications of global positioning systems (GPS), Technical report TR-24, The Ohio State University (1996).

[14] Recommendation ITU-T Y. 3031 Identification framework in future networks

[15] ITU-T Y.2057 (2011), Framework of node identifier and locator separation in IPv6-based next generation networks

[16] ITU-T Y.2720 (2009), NGN identity management framework

[17] ITU-T Y.3031 (2012), Identification Framework in Future Networks

[18] IETF RFC 4140, Hierarchical Mobile Ipv6 Mobility Management (HMIPv6)

[19] IETF RFC 3403 (2002), The Naming Authority Pointer (NAPTR) DNS Resource Record

Research on Blockchain Availability Modeling in P2P Network

Wang Zan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 18309220452@163.com

Zhong Lianjiong
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zhongli5180@sina.com

Fu Yanfang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: fuyanfang3000@aliyun.com

Dai Fei
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 8496579@qq.com

***Abstract*—Blockchain establishes reliable trust among parties that do not know each other and achieves credible data sharing and point-to-point value transmission in an epoch-making way. The requirement of the availability of block chain network becomes more and more important due to the dynamic and changeable characteristics of block chain P2P network. Therefore, for the characteristics of block chain P2P network system, this paper constructs a availability model based on Markov stochastic process theory, analyzes the steady-state availability and instantaneous availability of the model, and finally carries out experimental verification through simulation, hoping to provide beneficial inspiration and guidance for future research on block chain network .**

Keywords—Blockchain; P2p; Markov; Availability

I. INTRODUCTION

Block chain is a distributed system composed of encryption mechanism, storage mechanism, consensus mechanism, which can realize the peer-to-peer trading function of mutual trust without central server. The biggest feature of blockchain is decentralization and distribution, and the consensus mechanism of blockchain enables participating nodes to jointly provide services for the system and realizes similar functions of financial intermediaries in the central system. As shown in Figure 1, blockchain system can achieve fast synchronization of data without central server and consistency of consensus mechanism through P2P network. P2P realizes the sharing of data resources and cooperation of services among terminal devices by means of direct exchange. However the dynamic and changeable of blockchain P2P network, the requirement of the availability of block chain network becomes more and more important.

From the development trend of p2p network, the demand of modeling and verification evaluation on the availability of p2p network is more and more common and urgent. The establishment of high availability

DOI: 10.21307/ijanmc-2020-006

network is the basis of accurate and timely information exchange, so as to meet the demand of network system to provide high reliable services for various users.

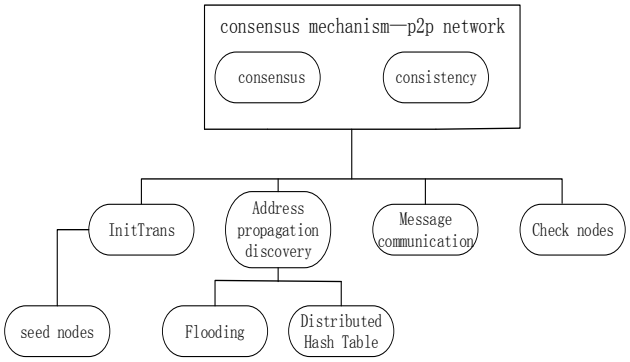


Figure 1. The architecture of p2p networks in the blockchain

II. RELATED WORK

The traditional concept of availability is a typical measure of reliability in reliability theory. It is an important parameter in reliability engineering that combines maintainability to represent the effectiveness of the system. "Reliability" can only reflect the probability of failure of the network system or components, while "availability" can reflect the quality of the network by considering the repairability of the network. Therefore, analyzing the availability of tact network system is an important index to evaluate the design of system networking, system stability and maintenance ability.

At present, availability studies mainly focus on the reliability and maintainability of the engineering capability of complex systems, and also on the availability of mission capability. Lianhong Zhou established the availability model of optical fiber communication system by using the state transfer equation[1]. Hailin Feng used Markov theory to study the steady-state availability of repairable network system, and analyzed the fuzzy availability of repairable network system and continuous kn(F) network. Fenghua Xie studied the availability of "N+X" UPS system and pointed out that the availability of parallel system was positively correlated with the number of redundant modules[2]. Jingle Zhang proposed a workflow availability modeling method based on random Petri net and studied and established the SPN model of e-commerce system[3]. Jikun Yang fused sample data with prior information, used fuzzy variables to deal with the uncertainty of sample data, and proposed a usability analysis method of navigation weapon system based on fuzzy Bayes[4].

Above all of these, we can find that the research on the availability of p2p network is relatively few and not very mature now. Therefore, research available technology based on P2P distributed collaborative network under different environmental conditions, analyze the unified modeling and expression of information, build a new generation of block chain network usability evaluation model, improve the information interaction and collaboration ability of block chain platform, and ensure the performance of system service and the formation of stable and reliable operation ability through availability technology.

III. DISTRIBUTED P2P NETWORK STRUCTURE BASEDON BLOCK CHAIN

Block chain technology is the first kind that can be globally distributed deployment of consensus agreement. Block chain system achieves efficient consensus through simple unauthenticated broadcast channel and block chain length competition mechanism. The typical blockchain system consists of network layer, consensus layer, data layer, intelligent contract layer and application layer[5]. The network layer is usually based on P2P protocol for communication between nodes; the consensus layer implements the consensus protocol, which can be freely selected according to the actual scenario, such as PoW based on workload proof, and PoS based on equity proof; the data layer is the data structure of the block chain, whose structural design is usually closely coupled with the application scenario according to the actual needs, each computing node is responsible for maintaining its own storage system; the intelligent

contract layer can perform different operations for different data input, this process is automatically executed based on the code, and consensus is reached across the network; the application layer is the basic business of the block chain system, such as financial services and data traceability, etc.

There is no central node in the block chain network based p2p, and any two nodes can be directly traded, each node is free to join and exit at any time. Therefore, the block chain platform usually chooses the P2P protocol which is completely distributed and can tolerate single point of failure as the network transmission protocol. Block chain network nodes have the characteristics of equality, autonomy and distribution, presenting a flat topology without any centralized authority nodes or hierarchical structure (as shown in Figure 2). Each node has such functions as route discovery, broadcast transaction, and broadcast block and find new nodes[6]. In the blockchain network, the P2P protocol of the blockchain network is mainly used for the transfer of transaction data and block data between nodes. The node listens to the data broadcast in the network all the time. It can only be processed after by verifying the validity of these transactions and blocks when receives new transactions and new blocks sent by the neighbor node.

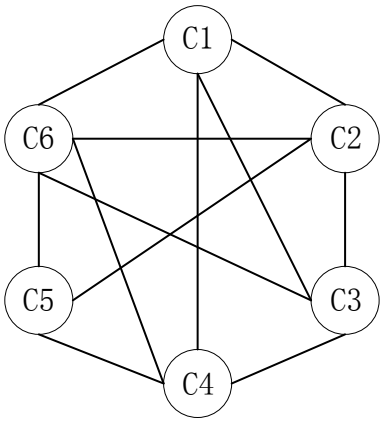


Figure 2. P2P network topology

IV. AVAILABILITY DEFINITION

Reliability is defined as the ability of a product to perform specified functions under specified conditions and within specified time periods. The higher the reliability, the lower the failure rate of the product. The simplest expression for reliability can be expressed as an exponential distribution, which expresses random failure.

$$R = e^{-\lambda t} \quad (1)$$

Where, t is mission time，λ is failure rate.

Availability is defined as the ability to be in a state of executable function under specified conditions and at specified times or time intervals, provided that the required external resources are guaranteed. It is the comprehensive reflection of product reliability, maintainability and maintenance guarantee. The formula is as follows:

$$\mathrm{A} = \frac{\mathrm{MTBF}}{\mathrm{MTBF+MTTR}} \quad (2)$$

MTBF(Mean Time Between Failure)refers to the average working time between two adjacent failures and is a reliability indicator of a product.

MTTR(Mean Time To Repair)describes the average repair time when a product changes from a failure state to a working state. In engineering, MTTR is a measure of the maintainability of a product, which is common in maintenance contracts and is used as a measure of service charges.

According to the above analysis, "reliability" only reflects the probability of failure of the network system or components, while "availability" considers the repairability of the network and better reflects the quality of the network. Steady-state availability and instantaneous availability are two characteristic quantities that reflect availability. Therefore, analyzing the steady state availability and instantaneous availability of blockchain P2P network system is an

important index to evaluate the design of system networking, system stability and maintenance ability.

V. MARKOVAVAILABILITY MODELING

P2P network system is a complex system, its nodes states are changing at anytime, while the factors causing the change of the state of nodes mainly include hardware and software errors, human errors, natural disasters, malicious attacked, causing serious consequences to the network system, and even causing the entire network paralysis. The probability of occurrence of the first several factors is relatively small, while as an artificial means, the probability of occurrence of malicious attack is very high in the real war environment.

A. Markov Chain Model of the System

In this paper, the P2P network system targeted is a multi-state Markov repairable system, assuming that its failures are caused by malicious attacks. The system is composed of several network nodes and several repair equipment. The life distribution of each node is $1-e^{-\lambda t}, t \geq 0, \lambda > 0$, and the repair time after node failure is $1-e^{-\mu t}, t \geq 0, \mu > 0$.All of these random variables are independent of each other, and after the fault node is repaired, its life will be the same as the new node. Multi-state means that the nodes of the network system are damaged by attack, which may break one, two or more nodes; the number of fault nodes is different means the system state is different.

In the process of system design, it is usually necessary to chose availability model to describe the availability of the system. The availability model adopted in this paper is the voting system model of take k in n. There are two types of voting system models of take k in n: one is a system of take k good nodes in n, which requires k or more of the n nodes of the system to be normal in order for the system to work normally, which is denoted as k/n[G]. The other kind is the system of take k bad nodes in n, which means that the system cannot work properly if k or more of the n nodes of the system fail, which is denoted as k/n[F][7].This paper chose the second system model. When the fault nodes are less than k, the system works normally and repairs the fault nodes. When there are k node failures, the system will fail. At this time, the remaining normal working nodes will also stop working and will not fail. The system will work again until one node under repair is repaired and there are less than k nodes under repair.

The system studied in this paper simplifies the actual situation, the n nodes of the system are considered to be one type of node. For example, because the importance of each node is different in practice, the life distribution and repair time of the node may be different, so the node types are not all the same. Another example is that when a node fails, its workload must be borne by other nodes, thus accelerating the loss of other nodes. Therefore, these idealized conditions are temporarily listed as assumptions, called the basic type for analysis.

The number of fault nodes in the system is defined as the state of the system, i.e.$E = \{0,1, \dots \dots, n-k+1\}$, where the working states of the system are$W = \{0,1, \dots \dots, n-k\}$, obviously, the working quality (quality of service) of the different working nodes included in these working states is different, but it is not differentiated in the basic type analysis, and all of them are considered as normal working.

X(t)=j indicates that there are j node faults in the system at time t that need to be repaired, i.e. it is in the state of j, where $j = \{0,1, \dots \dots, n-k+1\}$. Node life and maintenance time obey the exponential distribution, according to the above, can prove to be time continuous multi-state time homogeneous Markov chain, state transition probability in Δt time is shown in the following Formula (3).

$\boldsymbol{P_j}$ represents the probability that the system is in state j at time t. This is the Markov model of P2P network multi-state repairable system. The state transition probability diagram of the system is shown

in Figure 3: (the state of transition to itself is not shown in the figure)

$$\begin{cases} P_{j,j+1}(\Delta t) = (n-j)\Delta t + o(\Delta t),\ j = 0,1,\Lambda\ ,n-k \\ P_{j,j-1}(\Delta t) = j\mu\Delta t + o(\Delta t),\ j = 1,2,\Lambda\ ,n-k+1 \\ P_{j,j+1}(\Delta t) = -(n-j)\lambda\Delta t - j\mu\Delta t + o(\Delta t),\ j = 1,2,\Lambda\ ,n-k \\ P_{n-k+1,n-k+1}(\Delta t) = -(n-k+1)\mu \\ P_{j,k}(\Delta t) = o(\Delta t) \end{cases} \quad (3)$$

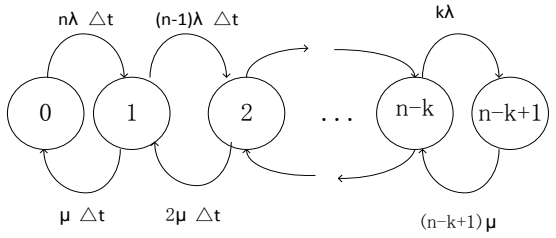


Figure 3. State transition diagram

The state transition probability matrix is represented by **A**, as follow:

$$\begin{bmatrix} -n\lambda & n\lambda & & & & & \\ \mu & -(n-1)\lambda-\mu & & & & & \\ & 2\mu & -(n-2)\lambda-2\mu & & & & \\ & & \Lambda & & & & \\ & & & (n-k)\mu & -k\lambda-(n-k)\mu & k\lambda & \\ & & & & (n-k+1)\mu & -(n-k+1)\mu & \end{bmatrix} \quad (4)$$

B. Steady-State Availability of the System

Steady-state availability is one of the reliability characteristics of markov repairable system, which means the proportion of the whole running time of the network without dismemberment when the network reaches the steady-state, it is the probability of the network being connected. This index is essentially the probability of connectivity at any time in the steady state. According to the properties of homogeneous Markov process:

$$lim_{t\to\infty} P_j'(t) = 0, \lim_{t\to\infty} P_j(t) = \pi_j \quad (5)$$

According to formula (5), we get

$$\boldsymbol{\pi A} = 0, \sum \pi_j = 1 \quad (6)$$

Where π is the probability distribution vector of state in steady state, $\pi = (\pi_0, \pi_1, \dots\dots, \pi_{n-k+1})$, Matrix **A** is substituted into Equation (6), and we get

$$\begin{cases} -n\lambda\,\pi_0 + \mu\,\pi_0 = 0 \\ n\lambda\,\pi_0 - ((n-1)\lambda+\mu)\,\pi_1 + 2\mu\,\pi_2 = 0 \\ n\lambda\,\pi_1 - ((n-2)\lambda+2\mu)\,\pi_2 + 3\mu\,\pi_3 = 0 \\ \Lambda\ \Lambda \\ (n-j)\lambda\,\pi_j - ((n-j+1)\lambda+(j+1)\mu)\,\pi_{j+1} + (j+2)\mu\,\pi_{j+2} = 0 \\ \Lambda\ \Lambda \\ k\lambda\,\pi_{n-k} - (n-k+1)\mu\,\pi_{n-k+1} = 0 \end{cases} \quad (7)$$

The value π_i can be obtained by solving the linear equations, and the steady-state availability of the system is

$$\boldsymbol{A = \sum \pi_i, i = 0, 1, 2, \dots\dots, n-k} \quad (8)$$

C. Instantaneous availability of the system

Instantaneous availability is also one of the reliability characteristics of markov repairable system. For some high-reliability systems like P2P networks, it takes a long time to reach their stable state, and the steady-state availability may also be disturbed during the system operation. Therefore, it is not enough to calculate the steady-state availability, but to consider the instantaneous index of the system.

From the C-K equation, we can obtain that the instantaneous probability is:

$$\begin{cases} P_{ij}(S+t) = \sum P_{ir}(s) P_{rj}(t) \\ P_0(0) = 1, P_1(0) = 0, \Lambda\ \Lambda\ , P_k(0) = 0 \end{cases} \quad (9)$$

The intuitive meaning of the C-K equation is start from state i and arrive at state j after s+ttime, and must first arrive at any state r after s time, and then transfer

from state r to state j after t time.

$P_0(0) = 1, P_1(0) = 0, \ldots\ldots, P_k(0) = 0,$ means that π_0 has a probability of 1 and the other states have a probability of 0 at the initial moment. Equation (9) is expressed as a matrix differential equation in the following form:

$$\begin{cases} \boldsymbol{P'(t) = P(t)A} \\ \boldsymbol{initial\ conditions\ P(0)} \end{cases} \quad (10)$$

$P(t) = P_0(t), P_1(t), \ldots\ldots, P_{n-k+1}(t)$ represents the distributed probability of the system in various states at time t, is the row vector, and is also a one-dimensional matrix. $\boldsymbol{P'(t)}$ represents the derivative matrix of instantaneous probability of the system at time t

This is a system of first order linear differential equations, the general form of its solution is:

$$P(t) = P(0)e^{At} = P(0)\sum_{n=0}^{\infty}\frac{t^n}{n!}A^n \quad (11)$$

Since **P** and **A** are matrices, it is difficult to find the analytic expression of the above equation, we can use the Laplace transform to find A simpler expression, and let

$$\boldsymbol{P^*}_j(s) = \int_0^{\infty} \boldsymbol{P}_j(t)\, e^{-st} dt, s > 0,$$

The Laplace transform of Formula (10) simplifies to

$$P^*(s) = \int_0^{\infty} P_j(t)\, e^{-st} dt = P(0)(SI - A)^{-1}, s > 0 \quad (12)$$

Where **I** is the identity matrix, $\boldsymbol{P(t)}$ can be obtained by inverting transform $\boldsymbol{P^*(s)}$, then the instantaneous availability of the system is

$$A(t) = \sum_{j \subseteq W} P_j(t) \quad (13)$$

However, it is difficult to obtain analytic expressions by $\boldsymbol{P^*(s)}$ inverse transformation. Results can be obtained by means of calculation tools, such as Matlab, Maple and other scientific calculation software.

VI. EXAMPLES VALIDATE

Take Figure 2 as an example, the number of network nodes is 6, and the minimum number of network nodes that can work normally is 4, that is4/6[F] voting system. According to Equation (7), steady-state availability π_i of each state is obtained, and then according to Equation (8), steady-state availability A of the network system is obtained. Table 1 is for steady-state availability of each states and the system when μ=0.5 hours and λ =2000 hours, Table 2 is for steady-state availability of each states and the system when μ=0.5 hours and λ =500 hours, and Table 3 is for steady-state availability of each states and the system when μ=1 hours and λ =100 hours.

TABLE I. STEADY-STATE AVAILABILITY CALCULATION RESULTS 1

	steady-state availability
π_0	0.9985013113
π_1	0.001497752142
π_2	0.0000009362705703
π_3	0.0000000002535257376
A	0.9999999997

TABLE II. STEADY-STATE AVAILABILITY CALCULATION RESULTS2

	steady-state availability
π_0	0.9880833998
π_1	0.01185708958
π_2	0.00005937422712
π_3	0.0000001284512973
A	0.9999998637

TABLE III. STEADY-STATE AVAILABILITY CALCULATION RESULTS 3

	steady-state availability
π_0	0.9420286502
π_1	0.05653222481
π_2	0.001423815141
π_3	0.00001530938621
A	0.9999846902

As can be seen from Table 1, 2 and 3, As can be seen from Table 1, 2 and 3, the probability of π_0 is the largest, and the probability of π_2 and π_3 are very small, even negligible.

Table 4 is the steady-state availability of a different μ and lambda system. Figure 1 is the curve trend diagram of Table 4, a more intuitive reflection of how steady-state availability varies with μ and λ

TABLE IV. STEADY-STATE AVAILABILITY CALCULATION RESULTSOF THE SYSTEM

	2000	500	100
2.5	0.9999999609	0.999996172	0.9988771093
1.5	0.9999999928	0.9999994802	0.9999020658
1	0.9999999979	0.9999998637	0.9999846902
0.5	0.9999999997	0.9999999837	0.9999979543

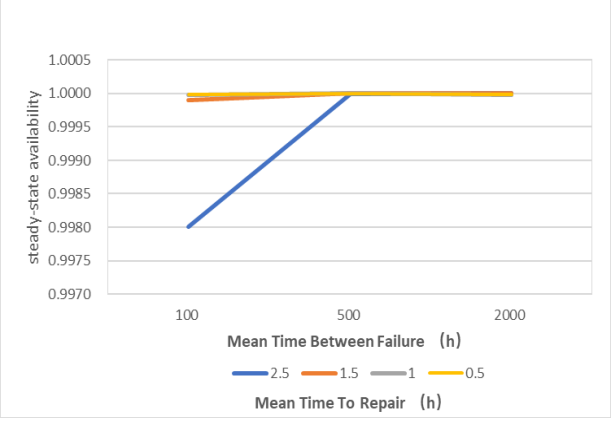


Figure 4. Steady-state availability trend diagram

In Table 4, the upper parameter is the node average failure parameter λ, the left parameter is μ. The x-coordinate of Figure 4 is the average node failure parameter λ, the y-coordinate is steady-state availability, and the four different curves are the average node repair time μ. As can be seen from Table 4 and Figure 4, under the Markov model, the steady-state availability of the system is very high, close to 1. The larger the λ, the smaller the μ, the more stable the system can be, which is consistent with the changing rules of P2P networks.

According to the analysis of 5.3, P2P network availability cannot be fully reflected by calculating steady-state availability. Therefore, we also need to calculate the instantaneous availability of the system, and calculate the instantaneous availability of the system according to Equations (12) and (13) with the parameters used in Table 1. The results are shown in the following Figure 5.

In Figure 5, the horizontal axis is time and the vertical axis is instantaneous availability. Since k=4, there are four curves in the instantaneous curves of multiple states, corresponding toπ_0, π_1, π_2 and π_3 respectively. In order to make the graph more intuitive to reflect the change of the curve, the graph is divided into two parts, the left one are π_1, π_2 and π_3, with the vertical coordinate of 0 to 0.004, and the right one isπ_0, with the vertical coordinate of 0.996 to 1. When four nodes fail, the entire network system is down and the network is unavailable. From the figure, we can see that π_0 i.e. the zero nodes that are damaged gradually decrease from probability 1, and become stable after reaching a certain time point. The instantaneous availability of π_0 is very high, which is close to 1 indefinitely. While the instantaneous availability of other states is relatively small, π_1 gradually increases with time until it becomes stable, while the instantaneous availability of π_2 and π_3 are very small, approaching 0 infinitely. It can be seen that the probability of system failure is very

small.

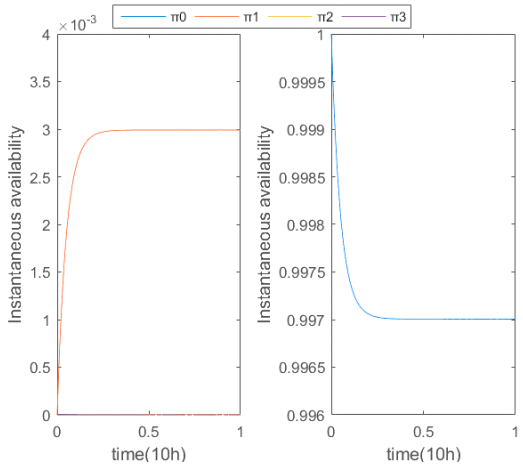


Figure 5. Instantaneous availability graphs of multiple states

VII. CONCLUSION

For P2P networks, the discrete time Markov chain method is used to establish a model of state transition in P2P networks in this paper, and the repairability of network nodes is considered. The calculation formulas of steady-state availability and instantaneous availability are given by using the model, and the results of the above calculation are obtained. The results are compared with the changing rules of the network system, which conforms to the availability requirements of the P2P network. It shows that this model has certain reference value in the availability test of tactical P2P network.

REFERENCES

[1] Lianhong Zhou. Availability of fiber optic systems[J].Tsinghua Univ (Sci& Tech).2002.01

[2] Fenghua Xie. Modular UPS reliability calculation and analysis[J].Fuzzy Systems & Mathematics, 2002.

[3] Jingle Zhang. Availability Analysis of Flexible Workflow Based on Stochastic Petri Net[J].Computer Engineering.2010.11

[4] Jikun Yang. Reliability and Availability Analysis of Missile Weapon System Based on Fuzzy Bayesian Theory[J]. Journal of Naval Aeronautical and Astronautical.2013

[5] Qifeng Shao. Block chain: Architecture and Research Progress[J].Chinese Journal of Computers. 2018,41(5):769-788

[6] Antonopoulos A M. Mastering Bitcoin : Unlocking Digital Cryptocurrencies[M].California : O'Reilly Media, 2014 : 137-172.

[7] Yongnian Wang. Reliability evaluation of k/n[G] system Availability Analysis of Flexible Workflow Based on Stochastic Petri Net. reliability analysis.2015:1-3

[8] Geng Yan. Markov modeling method for availability of twoitem system under passivation[J].Chinese Journal of Scientific Instrument .2016:1-6

[9] Yinan Jiang. The method of network reliability and availability simulation based on Monte Carlo[J]. ICQR2MSE.2012:1-5

[10] Patrick Wüchner. Queueing Networks & Markov Chains[M].Wiley-Interscience.1998:104-108

[11] Du Haidong. Simulation Assessment of Degradation System Reliability under Imperfect Maintenance[J]. Journal of System Simulation.2018,30(8):2834-2840

[12] Ivanovitch, Silva, Luiz Affonso, Guedes, Paulo, Portugal, et al. Reliability and availability evaluation of Wireless Sensor Networks for industrial applications[J]. Sensors, 2012, 12(1):806-838.

[13] Datta A. An overview of codes tailor-made for better repairability in networked distributed storage systems[J]. AcmSigact News, 2013, 44(1):89-105.

[14] YUAN Yong. Blockchain: The State of the Art and Future Trends [J]. ACTAAUTOMATICASINICA, 2016, 42(4)481-494

[15] Wenli Zhou. Survey of P2P technologies[J]. Computers Engineering and Design , 2006, 27(1):76-79

[16] Xuelong Wang, Jing Zhang,, et al. Survey on peer-to-peer key technologies[J]. Application Research of Computers, 2010, 27(3):801-805,823.

Research on the Tunnel Geological Radar Image Flaw Detection Based on CNN

Li He
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1003294436@qq.com

Wang Yubian
Department of Railway Transportation Control
Belarusian State University of Transport
34, Kirova street, Gomel, 246653
Republic of Belarus
E-mail: alika_wang@mail.ru

***Abstract*—Tunnel geological radar image has been widely used in tunnel engineering quality detection for its advantages of fast, nondestructive, continuous detection, real-time imaging, intuitive data processing and high detection accuracy. However, the traditional defect detection method, which is judged by surveyors visually, consumes energy. In order to detect the quality of tunnel engineering accurately and quickly, an improved method of void defect detection based on Faster RCNN (Regional Convolutional Neural Network) is proposed in depth learning. The image data of the tunnel geological radar is collected for annotation, which fills the blank of the defect data set in the tunnel engineering. Through the method of this paper proposed, the feature extraction is optimized to improve the performance of the detection model, and the detection accuracy of the model is verified by expert knowledge.**

Keywords-Tunnel Geological; Radar Image; Flaw Detection; CNN

I. INTRODUCTION

There are a large number of tunnel projects in the construction project. The quality defects of the tunnel may affect the construction schedule, increase the engineering cost, damage the mechanical equipment and even endanger the lives of the constructor. Geological radar detection method[1-2] is the mainstream method of tunnel lining detection at present, and has excellent performance in the detection of reinforcement and arch spacing, plain concrete structure, etc.[3]. However, the traditional survey situation is that the site construction surveyors scan the survey images generated by radar equipment one by one according to their expert knowledge. This traditional method has a large workload, a large human factor, and a certain rate of omission and error.

In recent years, with the continuous improvement of GPU, the field of deep learning is booming. In 2006, Hinton[4] and other researcher proposed the concept of deep learning, using convolutional neural network (CNN) to learn features from data. In 2012, in the ImageNet image classification competition, Alex Krizhevsky's team proposed the deep convolutional neural network AlexNet for the first time. AlexNet won the champion with an accuracy rate of 15.3% higher than the second place, which made people have a further understanding of the application of convolutional neural network in visual tasks. Girshick R proposed the regional convolutional neural network (RCNN) model, which USES the Selective Search method to select candidate regions and USES multiple

support vector machines to classify features, thus achieving target detection.

In 2015, Girshick R proposed Fast RCNN[6], which is an improved version of RCNN and adopts RoI Pooling to share the parts with a large amount of calculation to improve the working efficiency of the whole model. Later, Ren improved again on the basis of the original network, introduced RPN layer, and designed a Faster RCNN model[7], aiming at the problem that the running time of extracting candidate feature regions was slow, which achieved good results. Compared with artificially designed features, features extracted by convolutional neural network have better robustness and stronger semantic information, and great achievements have been made in computer vision fields such as face recognition[8-9], target detection, and speech recognition[10-11].

This paper selects Faster RCNN network as the basic algorithm framework for tunnel GPR (Ground Penetrating Radar) detection. The framework of Faster RCNN network is introduced. However, if the original Faster RCNN model is directly applied to the tunnel GPR detection in the actual scene, there may be two disadvantages:

1) The data sets collected on site for training are relatively small, which may lead to incomplete learning of the learning model and easy over fitting[12].

2) There are many interference factors in the tunnel, resulting in complex image features of the defect. At the same time, the radar images are all manually experienced by field surveyors, and there is no uniform standard, so the sharpness is quite different. It will cause RPN to produce more negative sample space, and the network model is difficult to converge[13].

Based on the above reasons, this paper proposed Faster - RCNN model to expand original data sets, based on the data increase[14] and combining with GA – RPN[15] for an improvement in the target detection and evaluation index in GIoU on border regression optimization[16], in order to overcome the above shortcomings, further improve the accuracy of tunnel geological radar image detection.

II. KEY TECHNOLOGIES AND EQUIPMENT

As a relatively mature geophysical prospecting method, geological radar method has the advantages of high resolution, fast detection speed, non-destructive and radar image visualization, etc., and has become the most important method for tunnel lining quality detection. There are obvious abnormal reactions to the defects of tunnel lining, such as local un-compactness, voidage, insufficient thickness and lack of reinforcement.

A. Technical principle

Different defects have different reflections in radar images. The electromagnetic waves emitted by GPR will generate reflection and refraction on the surface of the medium with different dielectric constants[17]. The dielectric constants of common materials are shown in table 1 below.

TABLE I. DIELECTRIC CONSTANTS OF COMMON MATERIALS

Material	Dielectric constant	Velocity (mm/ns)
atmosphere	1	300
water	81	30
concrete	5-8	55-120
Sand (dry)	3-6	120-170
Sand (wet)	25-30	55-60
pitch	3-5	134-173

Reflection and refraction conform to the law of reflection and refraction. The energy of reflected wave and refracted wave depends on the reflection coefficient R and refracted coefficient T:

$$R = \frac{\sqrt{\varepsilon_1} - \sqrt{\varepsilon_2}}{\sqrt{\varepsilon_1} + \sqrt{\varepsilon_2}} \quad (1)$$

In the above equation, ε1 and ε2 are the relative permittivity of the upper and lower media at the interface respectively. According to Formula (1), when the electromagnetic wave propagates to the interface with dielectric constant difference, the reflected electromagnetic wave energy will change, which is reflected in the radar image as positive and anti-peak anomalies. The defects in lining such as local un-compaction, voidage, insufficient thickness and lack of reinforcement have obvious dielectric differences with concrete, which provide a good geophysical foundation for the application of GPR.

B. Equipment

The equipment of detection equipment of GPR system in the project is shown in Figure 1.

Figure 1. GPR system detection equipment

Geological ground penetrating radar method is the use of high frequency electromagnetic wave transmitting antenna will be in the form of a pulse in the concrete surface emission to concrete, the concrete interface reflection and defects return to the surface, by the receiving antenna to receive the echo signal, the treatment of radar system, radar image, through the analysis of the radar image processing, the interpretation of the data on the basis of this, so as to achieve quality nondestructive testing of lining. Its detection principle is shown in Figure 2.

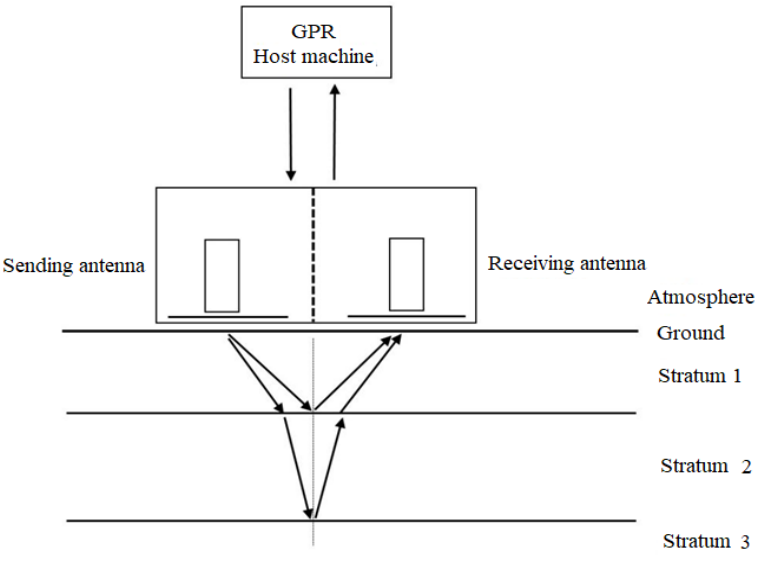


Figure 2. Principles of geological radar detection

C. The Radar imaging profile

Profiles of GPR images are usually recorded in the form of pulse reflected waves, or in the form of gray or color profiles. In this way, the in-phase axis or the iso-gray line can be used to represent the underground reflector or the target body. On the waveform recording chart, waveforms are recorded in the vertical direction of the survey line at all measurement points, forming the radar imaging profile.

In the tunnel lining detection, the voids mainly appear between the inner lining, the second lining and the surface layer. Due to the large dielectric difference between the void and surrounding media, when the electromagnetic wave propagates between concrete and atmosphere, air and surrounding rock, it will generate two strong reflections at the upper and lower interfaces.

Because the electromagnetic wave in the air attenuation is small, and the concrete and air dielectric difference is large, so in the void to produce a strong multiple reflections, electromagnetic wave in the air medium propagation frequency is relatively high.

At the upper interface of voids, the electromagnetic wave goes from concrete to the air medium. According to the law of reflection (Formula 1), the reflection coefficient is positive, showing a negative wave peak. At the gap interface, the reflection coefficient of electromagnetic wave from air to concrete medium is negative, showing a positive wave peak. The picture is as shown in Figure 3.

In the image, white is the strongest color of positive reflection and black is the strongest color of negative reflection. In theory, there is a gap defect and the radar image shows two sets of reflected signals. In the actual survey, the second reflection signal may be lost due to signal interference.

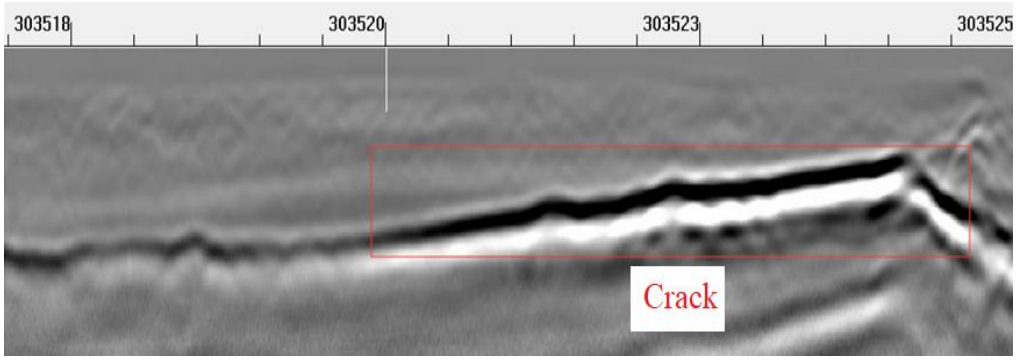


Figure 3. GPR image of tunnel ejection

III. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNN) can use local operations to abstract the representations in a hierarchical way in image recognition [18]. Two key design ideas have driven the success of convolution architecture in computer vision. First, CNN USES the 2D structure of the image, and the pixels in adjacent areas are usually highly correlated. Therefore, instead of using one-to-one connections between all pixel units (as most neural networks do), CNN can use local connections in packets. Second, the CNN architecture relies on feature sharing, so each channel (that is, the output feature graph) is generated by convolution with the same filter at all locations.

Deep learning methods of convolutional neural network are mainly divided into single-stage (eg.SDD, YOLO) and two-stage (eg.RCNN series). Single-stage generates detections directly on the picture through

calculation. Two-stage extracts the proposal first, and then makes the second amendment based on the proposal. Both are relatively single-stage fast, low precision. In this paper, it is proposed to use two-stage Faster RCNN because of the high accuracy of two-stage[19].

This paper adopts Faster RCNN as the basic framework, as shown in Figure 4. In fact, Faster RCNN can be divided into four main contents:

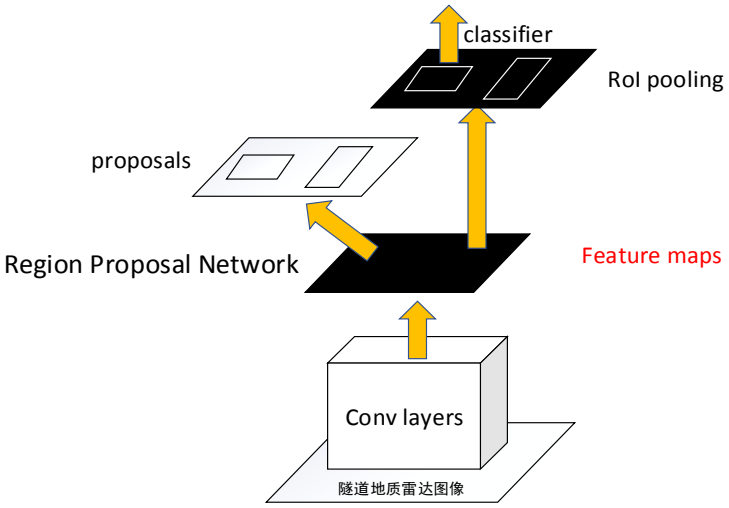


Figure 4. Structure chart Faster RCNN

1) Convolutional layers. It is used to extract the features of un-blemishes on the GPR image. The input is the whole image and the output is the extracted features, which are called feature maps.

2) Region Proposal Network. It is used to recommend candidate regions; this network is used in place of the previous search selective. The input is an image (because the RPN network and Fast RCNN share the same CNN here, so the input can also be considered feature maps), and the output is multiple candidate areas to filter out gaps in features and = perform a preliminary border regression. It shares the convolution feature of the whole graph with the detection network, solves the speed bottleneck of the original selective search method, and greatly improves the speed of target detection.

3) RoI pooling. Its function is input different sizes and converted into output of fixed length, and the input and output are the same as RoI pooling in Faster RCNN.

4) Classification and regression. The output of this layer is the class of the fully connected neural network of candidate region, and the exact location of the candidate region in the image.

The whole process is shown in Figure 5.

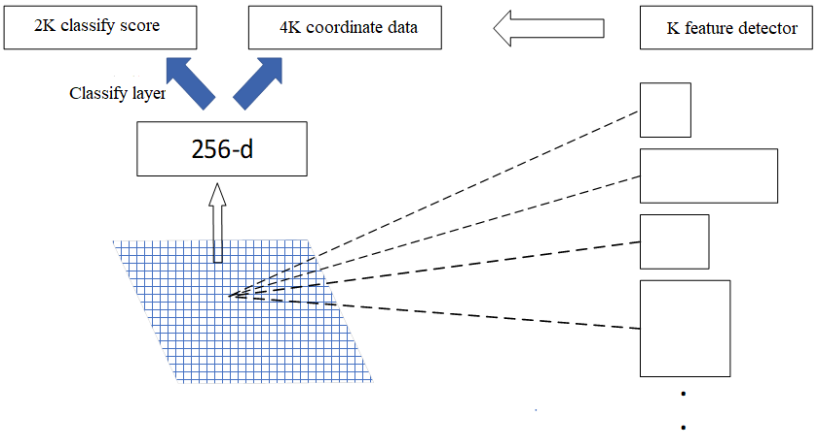


Figure 5. Structure chart of RPN

IV. GENERATING ANCHOR -RPN

Regional Anchor is an important cornerstone of target detection method at present. Most good target detection algorithms rely on the Anchors mechanism to evenly sample a given location in space with predefined sizes. There are two major problems in the process of generating Anchor by the traditional two-stage anchor-based method:

1) Inefficient is low. The existing method makes sliding Windows on the feature map and generates thousands of anchors. However, there are only a few objects in one picture, which leads too many negative samples.

2) Unreasonable a priori assumptions. When generating Anchor, it is assumed that the scale of Anchor or the length-width ratio is several fixed values. These values tend to change and change with the dataset.

According to the above Hamid Rezatofighi proposed the guided anchor in 2019. The Guided anchor mechanism works as follows: the position and shape of the target can be represented by (x, y, w, h). (x, y) represents the coordinates of the object's position in space.

If the box of the object is drawn on a given input picture, the following distribution can be obtained:

$$p(x, y, w, h \mid I) = p(x, y \mid I)p(w, h \mid x, y, I) \quad (2)$$

There are two important information can be obtained from the above equation: (1) the specific region of the object in the image; (2) the size and proportion of the object are closely related to its location. The anchor generated model is therefore designed to contain two branches: one for positioning and one for shape prediction.

A. *Anchor generation modules.*

That is position prediction module. The goal is to predict which regions should be used as the center points to generate the anchor, which is a dichotomy problem, to predict whether or not the object is the center. Two branches were added to predict the confidence of each pixel (corresponding receptive field) on the feature graph, as well as the corresponding width and height.

A target is considered a target if its confidence is greater than a specific domain value. Obviously, the process of obtaining this proposal is different from that of sliding window, which can reduce a large number of negative samples (only one proposal can be generated by making more pixels on each Feature map).

In addition, since the width and height are also regressed by CNN, there is no scale for the object, and

the width and height are compared to any prior assumptions.

B. Feature adaption modules.

This module actually quotes from the idea of deformable convolution. Firstly, the width and height of each point can be obtained by using the Anchor generation module. The width and height are represented by a 2-channel characteristic graph, and then the offset field is obtained by convolving again on the 2-channel characteristic graph. Finally, the offset field is applied to the characteristic graph.

V. IMPROVE IOU ALGORITHM

In the target detection, it depends on the regression of boundingbox coordinates to obtain accurate positioning effect. IoU (Intersection-over-Union) is an important concept in target detection. In the anchor-based method, it is not only used to determine positive samples and negative samples, but also to evaluate the distance between the predictbox and ground-truth, or the accuracy of the predictbox.

One of the better features about the IoU is that it's scale insensitive. In the regression task, the most direct indicator to judge the distance between predict box and gt is the IoU, but the loss used is not suitable. Since loss cannot reflect the regression effect, IoU can get different values according to different situations, which can most directly reflect the regression effect. But there are two problems with using IoU directly as the loss function:

A. *If the two boxes do not intersect, by definition, IoU=0, it does not reflect the distance between them.* At the same time, because loss=0, there is no gradient return, no learning and training.

B. *The IoU cannot accurately reflect the degree of coincidence between the two.* As shown in figure 6 below, IoU is equal in all three cases, but their coincidence degree is different. The graph on the left has the best regression effect, while the graph on the right has the worst.

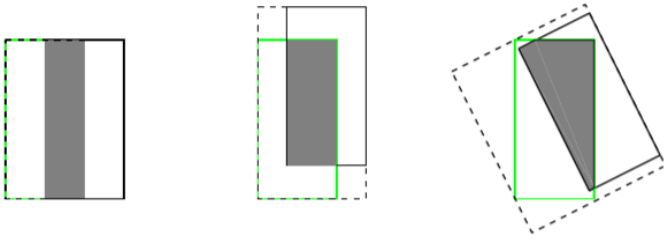


Figure 6. Same border regression of IoU

To solve the above problems, Rezatofighi, Hamid and other researcher proposed GIoU in 2019, and its algorithm formula for GIoU is as follows:

Algorithm 1: Generalized Intersection over Union

Input：Two Random image: A， $B \subseteq S \in R^n$

Output：GIoU

1. Find the smallest target graph C that surrounds A and B，C satisfies the condition: $C \subseteq S \in R^n$
2. $IoU = \frac{|A \cap B|}{|A \cup B|}$
3. $GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|}$

C. Characteristics of GIoU

1) Similar to IoU, GIoU is also a distance measure. As a loss function, L_{GIoU}=1-GIoU, which meets the basic requirements of the loss function.

2) GIoU is Insensitive to scale.

3) GIoU is the lower bound on IoU, and in the case of wireless overlap between the two boxes, IoU=GIoU.

4) The value of IoU is [0,1], but the value range of GIoU is symmetric [-1,1]. According to the above formula, it can be seen that GIoU is always less than or equal to IoU. In addition, for IoU, its range is [0,1], while the range of GIoU is [−1,1]. When the two shapes coincide completely, there is GIoU=IoU=1. When there is no overlap between the two shapes, IoU= 0, and the subtraction is -1.

5) Unlike the IoU, which only focuses on the overlapping area, the GIoU not only focuses on the overlapping area, but also focuses on other non-overlapping areas, which can better reflect the coincidence degree of the two. Since GIoU introduces C containing two shapes of A and B, it can still be optimized when A and B do not coincide.

VI. Experimental Methods

A. Data to enhance

Compared with traditional images, the number of GDAR images that can provide deep learning training is generally less, and the lack of data will lead to over fitting of the model. In order to ensure the generalization ability and recognition effect of the model after training, it is more important to improve the training performance of the deep learning method by means of data augmentation. The small manually labeled data set was taken as the initial sample, and the initial sample set was rotated and compound rotated, and all the operation results obtained were taken as the training data set after data enlargement.

B. Training methods

After the data enhancement, the images of plain concrete and reinforced concrete were sent into the neural network to start the training model. Faster - RCNN model training method using alternate optimization method (alternating optimization), it is divided into four steps.

1) Stage1_rpn_train.pt

RPN network was trained separately, and the trained void model was initialized with the model of ImageNet, and the parameters were adjusted in an end to end manner. backbone+rpn+fast rcnn——>backbone1+rpn1+fast rcnn, backbone, rpn, parameter updating;

2) Stage1_fast_rcnn_train.pt

Fast RCNN is a separate training detection network. Proposals for training come from RPN net in step 1. The model initialization adopts the ImageNet model. backbone+rpn1+fast rcnn——>backbone2+rpn1+fast rcnn1, backbone, fast rcnn, parameter updating;

3) Stage2_rpn_train.pt

The RPN model was initialized with the parameters of the second step Fast Rcnn, but the convolutional layer was fixed during the training and only the parameters belonging to RPN were adjusted. backbone2+rpn1+fast rcnn1——>backbone2+rpn2+fast rcnn1, rpn, parameter updating;

4) Stage2_fast_rcnn_train.pt

Keep the Shared convolutional layer fixed; fine-tune the remaining parameters of Fast rcnn with the proposals of the RPN output adjusted in step 3 as input. Backbone2+rpn2+fast rcnn1→backbone2+rpn2 + fast rcnn2, fast rcnn, parameter updating.

The above four steps are shown in figure 7.

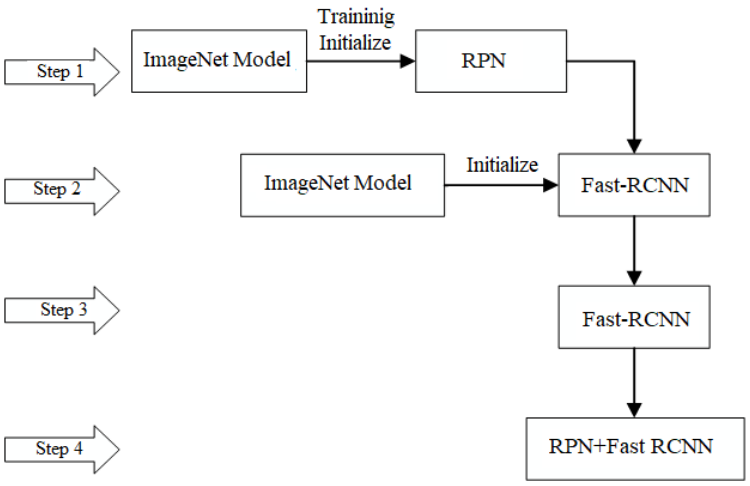


Figure 7. Network model training process

C. *Evaluation index*

The general formula for calculating the accuracy is: the number/total number of accurate classification ×100% [20]. This paper uses traditional evaluation criteria

Accuracy: ACC= (TP+TN)/ (TP+TN+FP+FN)

Precision: P= TP/ (TP+FP), the proportion of positive classes in the results after classification.

Recall: RECALL=TP/(TP+FN), the proportion of all positive examples divided into pairs.

TP means that the positive sample is correctly identified as a positive sample.

TN means that negative samples are correctly identified as negative samples.

FP indicates that a negative sample is incorrectly identified as a positive sample.

FN means that a positive sample is incorrectly identified as a negative sample.

The identification criteria for identification of void in the tunnel lining in the image are defined as the type and location of void. The probability size and the coincidence degree of the identification box and the marker box are determined as empty. Through the analysis of the sample situation, if the probability of the defect of the Geology radar image and the GIoU reach more than 50%, the gap will be recognized.

VII. CONCLUSION

In tunnel construction, geological radar is used to detect highway tunnel engineering, and the radar scanning result map can be used to realize advanced geological forecast, and the hidden danger of geological defects in the tunnel can be found. Based on the analysis of the elation principle, a method based on Faster RCNN is proposed in this paper to extract the position of the elation in the second lining. Compared with the traditional identification method, this method has the characteristics of Faster and higher accuracy. Since this data set is only plain concrete and reinforced concrete, the scale of annotated data will be further expanded to enrich the diversity of samples to improve the performance of the model.

REFERENCES

[1] Li Wendi. Analysis of GPR image features of tunnel lining defects. [J]. Fujian Building Materials, 2019(01):22-24.

[2] Zhang Chi. Research on detection of lining voids of reinforced concrete structures based on geological radar method [J]. Railway survey, 2018, 44 (03):35-38.

[3] Liu Jinlong, Tan hailiang. Application of geological radar in detecting soil defects [J]. Engineering quality, 2018, 36 (01):73-75.

[4] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18: 1527-1554.

[5] Krizhevsky A, Sutskeever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the Acm, 2012, 60(2): 1097-1105.

[6] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 1440-1448.

[7] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[8] Sun Y, Liang D, Wang X G, et al. DeepID3: face recognition with very deep neural networks [J]. Computer Science, 2015, 2(3): 1-5.

[9] Luiz G H, Robert S, Luiz S O. Written dependent feature learning for offine signature verification using deep convolutional neutral networks [J]. Pattern Recognition, 2017(70): 163-176.

[10] Abdelhamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(10): 1533-1545.

[11] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Seattle, WA, USA. New York: IEEE, 779-788.

[12] Du Yuhong, Dong Chao-qun, etc. Application of improved Faster RCNN model in cotton fiber identification [J/OL]. Advances in laser and optoelectronics:1-14[2019-11-25].

[13] Xu shoukun, Wang yaru, Gu yuwan etc. Research on the detection of helmet wearing based on improved FasterRCNN [J/OL]. Computer application research: 1-6[2019-11-25].

[14] Song Shang-ling,Yang Yang etc. Pulmonary nodules detection algorithm based on Faster-RCNN [J/OL]. Chinese journal of biomedical engineering: 1-8[2019-11-25].

[15] Wang J, Chen K, Yang S, et al. Region Proposal by Guided Anchoring [J]. 2019. R ezatofighi H, Tsoi N, Gwak J Y, et al. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression [J]. 2019.

[16] Rezatofighi H, Tsoi N, Gwak J Y, etc. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression[J]. 2019.

[17] Zheng Lifei, Xiao lito, Li Xiaoqing. Forward modeling and application of geological radar advance prediction [J]. Communications science and technology, 2018(2):76-81.

[18] Wu Zhengwen. Application of convolution neural network in image classification. Chengdu: University of Electronic Science and technology of China, 2015

[19] Lin Gang, Wang Bo etc. Multi-target detection and positioning of power line inspection image based on improved faster-RCNN [J]. Power automation equipment, 2019, 39(05):213-218.

[20] Sainath TN, Kingsbury B, Saon G, et al. Deep convolutional neural networks for large-scale speech tasks [J]. Neural Networks, 2015, 64:39-48.

Research on Control Strategy of Matrix Converter Motor System

Xu Shuping
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 563937848@qq.com

Su Xiaohui
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China

Chang Yichen
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China

Guo Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China

***Abstract*—In order to solve the problem of the industrial application of Matrix Converter (MC), the Matrix Converter Motor System (MCMS) is proposed. The control strategies of MCMS are studied based on the control of MC and the motor. The effects of voltage vector and direct torque control are analyzed. The simulation result show the disadvantages of usual means can be conquered by new means and it can obtain more high accuracy and robustness. The dynamic and static performances of direct torque control based on it are better than that of conventional direct toque control. The new ameliorated control strategy is proposed.**

Keywords-Matrix Converter; VVVF Control; Vector Control; Direct Torque Control

I. INTRODUCTION

Matrix converter output frequency is independent of input frequency limits, no DC link, small size, compact structure, good quality of input / output current, input power factor adjustable, and can be achieved two-way flow of energy. It is in line with the ideal standard of electric drive and expected to become the mainstream technology of future Electric transmission[1-3]. The biggest obstacle of hinder matrix converter access to industrial application is low voltage transfer. In the line shape modulation, the maximum of voltage transfer ratio of matrix converter is 0.866[4-5].

Therefore, scholars have proposed various solutions. Matrix converter motor system is a combination of both matrix converter and motor, so control strategy is also a combination of matrix converter and motor control method. This paper proposed the concept of matrix converter motor system aimed at the problems of matrix converter enter into industrial applications, and researched the control strategy of matrix converter motor system combined with matrix converter control method and the motor speed method.

II. MATRIX CONVERTER MOTOR SYSTEM

Matrix converter motor system is composed of matrix converter, matrix converter motor, input filter, clamp circuit and other components shown in Figure 1.

In which, the rated magnetic flux of Matrix converter motor design based on the maximum voltage transfer ratio of 0.866 of matrix converter to meet the requirements of matrix converter voltage transfer ratio and take into account the integration issues with the matrix converter.

In order to filter the high harmonic in input current caused by the switch frequency, matrix converter motor system needs to set the input filter. View from the power side of the matrix converter is a current source, so a LC second order filter circuit adopted, the design criteria of LC input filter as follows: cutoff frequency of input filter should be lower than the switch frequency; the power factor angle offset caused by the filter should be minimum under the minimum output power set; A certain voltage and electric resistance, minimize size or weight of the input filter by selecting capacitors with different power density; At rated

DOI: 10.21307/ijanmc-2020-008

current, the voltage caused by filter inductor should be drop the minimum to reduce adverse effects to voltage transfer ratio.

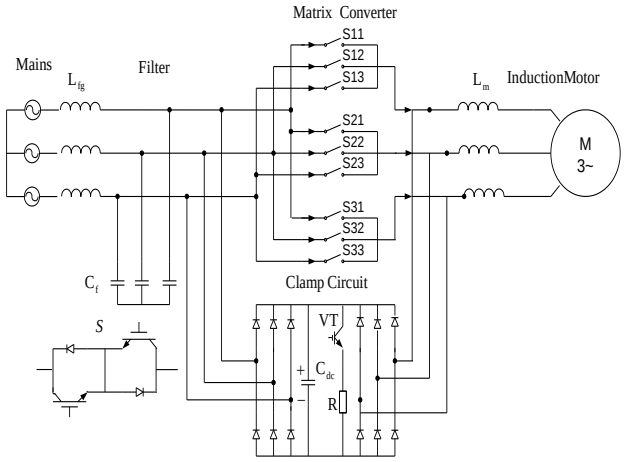


Figure 1. The topology of the matrix converter motor system

III. MATRIX CONVERTER CONTROL STRATEGIES AND MOTOR CONTROL STRATEGIES

Three-phase / three phase matrix converter consists of 9 two-way switches. Two basic principles of Matrix Converter Control Strategies as follows: Two input phase can not be connected meanwhile to an output phases to prevent short circuits; any output phase must ensure that an input is linked in order to prevent inductive load open circuit. Constraints in the above principles, matrix converter has a total of 27 switch states.

Matrix converter control strategy has switch function method, dual-voltage vector method, space vector modulation method and hysteresis current tracking method. Switching function method is calculated the duty cycle of 9 switches by the relationship between the actual inputs of inverter and the input requirements[6]. Dual voltage synthesis method is in each switching cycle, used linear combinations of the two input lines voltage to synthesize output line voltage of two phase law[7]. Space vector modulation method is based on the virtual DC link ideas, the matrix converter is equivalent to the dual PWM converter of abolition DC link, and used space vector modulation to both rectifier link and inverter link which equivalent AC - DC - AC structure, and then back to matrix converter[8]. Hysteresis current tracking method use hysteresis track on the output current to track the reference current[9].

Motor control strategy has VVVF control, vector control and direct torque control. Through the speed regulator, VVVF control can achieve closed loop of speed; Vector control to achieve double closed loop of rotate speed and torque component of stator current, in which rotate spend loop is the outer ring, torque component of stator current for the inner ring; Direct torque control achieve double-loop of speed and torque; where speed loop for the outer ring, the torque for the inner ring.

IV. MATHEMATICAL MODEL OF PMSM

For the vectors relationship between flux voltage and current in permanent magnet synchronous motor, d, q coordinate system fixed in the rotor coordinate system; x, y coordinate system fixed to the stator rotating coordinate system, stator flux $\varphi = \{\varphi_d, \varphi_q\}^T$ as state vector, stator current $i = \{i_d, i_q\}^T$ as output vector, $U = \{u_d, u_q, \varphi_f\}^T$ as the input vector. So the motor properties model can be rewritten in matrix form as follows:

$$\frac{d}{d_x}\begin{bmatrix}\varphi_d\\ \varphi_q\end{bmatrix} = \begin{bmatrix}-\frac{R_s}{L_d} & \omega_e\\ -\omega_e & -\frac{R_s}{L_q}\end{bmatrix}\begin{bmatrix}\varphi_d\\ \varphi_q\end{bmatrix} + \begin{bmatrix}1 & 0 & \frac{R_s}{L_d}\\ 0 & 1 & 0\end{bmatrix}\begin{bmatrix}u_d\\ u_q\\ \varphi_f\end{bmatrix}$$

$$\begin{bmatrix}i_d\\ i_q\end{bmatrix} = \begin{bmatrix}\frac{1}{L_d} & 0\\ 0 & \frac{1}{L_q}\end{bmatrix}\begin{bmatrix}\varphi_d\\ \varphi_q\end{bmatrix} + \begin{bmatrix}0 & 0 & -\frac{1}{L_d}\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}u_d\\ u_q\\ \varphi_f\end{bmatrix}$$

φ_d, φ_q are the axis flux; u_d, u_q are the axis voltage.

V. THE STATOR FLUX BASED ON FULL-DIMENSIONAL STATE OBSERVER

A. structure of stator flux full-dimensional state observer

In practice, there is not all state variables can be measured directly, so we can reconstruct state variables by variables can be measure in system to construct a system identical to the original artificially system. The system is constructed artificially, so its state variables can be measured completely.

Construction System:

$$\begin{cases}\dot{\hat{\psi}}_s = A\hat{\psi}_s + Bu_s\\ \hat{i}_s = C\hat{\psi}_s + Du_s\end{cases}$$

$\hat{\psi}_s$ is the state vector for estimated value of ψ_s which original system actual vector need to observe.

$$A=\begin{bmatrix} -\frac{R_s}{L_d} & \omega_e \\ -\omega_e & -\frac{R_s}{L_q} \end{bmatrix},\ B=\begin{bmatrix} 1 & 0 & \frac{R_s}{L_d} \\ 0 & 1 & 0 \end{bmatrix}$$

$$C=\begin{bmatrix} \frac{1}{L_d} & 0 \\ 0 & \frac{1}{L_q} \end{bmatrix},\ D=\begin{bmatrix} 0 & 0 & -\frac{1}{L_d} \\ 0 & 0 & 0 \end{bmatrix}$$

$$\hat{\psi}_s=\begin{bmatrix} \hat{\psi}_{sd} & \hat{\psi}_{sq} \end{bmatrix}^T, u_s=\begin{bmatrix} u_{sd} & u_{sq} & \psi_f \end{bmatrix}^T,$$

$$i_s=\begin{bmatrix} i_{sd} & i_{sq} \end{bmatrix}^T$$

so: $\dot{\psi}_s-\dot{\hat{\psi}}_s=A(\psi_s-\hat{\psi}_s)$

solution: $\psi_s-\hat{\psi}_s=e^{At}\left[\psi_s(0)-\hat{\psi}_s(0)\right]$

Known by the above equation: When $\psi_s(0)=\hat{\psi}_s(0)$, that $\hat{\psi}_s=\psi_s$. However, in general, to ensure initial conditions identical at any time is impossible. Therefore, in order to eliminate state error, we can introduce error feedback $(\psi_s-\hat{\psi}_s)$ according to the basic principles of control theory, that is, error feedback $(i_s-\hat{i}_s)$. The advantages of introduces feedback terms as follows: firstly, when the initial state ψ_s and $\hat{\psi}_s$ have a deviation, it can control the deviation not diffusion or shock with time increase; secondly, it can overcome the shortcomings of the open-loop observer can not be adjusted when matrix A changes that made the deviations of ψ_s and $\hat{\psi}_s$ more worse.

The state estimation error of observer can be obtained from Figure 1 is $\boldsymbol{\psi}_s-\hat{\boldsymbol{\psi}}_s=\mathrm{e}^{(\mathbf{A}-\mathbf{GC})t}[\boldsymbol{\psi}_s(0)-\hat{\boldsymbol{\psi}}_s(0)]$. Clearly, as long as the eigenvalues of coefficients $\mathbf{(A-GC)}$ of chosen observer all have negative real parts, it can make state estimated value $\hat{\boldsymbol{\psi}}_s$ incremental approach to the real value $\boldsymbol{\psi}_s$.

B. Pole Placement

The pole of observer is the eigenvalues of $(A-GC)$ which is important to observer performance. Usually pole placement of observer should make the response speed of observer faster than the system model. That can make the observation error fast convergence. To get the eigenvalues of $(A-GC)$, this paper used induction motor pole configuration method in documents, that is the method of pole placement of observer configured direct proportional to motor model.

Consider the poles of motor, that is, the eigenvalues of $|sI-A|=0$ that eigenvalues of the following formula:

$$s^2+(\frac{R_s}{L_d}+\frac{R_s}{L_q})s+\frac{R_s}{L_d}\cdot\frac{R_s}{L_q}+\omega_e^2=0$$

Two poles can be got:

$$s_1=\frac{-(\frac{R_s}{L_d}+\frac{R_s}{L_q})+\sqrt{(\frac{R_s}{L_d}+\frac{R_s}{L_q})^2-4(\frac{R_s}{L_d}\cdot\frac{R_s}{L_q}+\omega_e^2)}}{2}$$

$$s_2=\frac{-(\frac{R_s}{L_d}+\frac{R_s}{L_q})-\sqrt{(\frac{R_s}{L_d}+\frac{R_s}{L_q})^2-4(\frac{R_s}{L_d}\cdot\frac{R_s}{L_q}+\omega_e^2)}}{2}$$

$$(j=\sqrt{-1})$$

Set observer poles is k times than motor poles, that is, $OS_1=k\,S_1 \quad OS_2=k\,S_2$

The position of the motor poles and observer's poles k =1.2 in the article.

Using the powerful scientific computing advantages of Matlab and control system toolbox, according to the robust pole assignment algorithm and equation, the gain matrix of the observer **G** can be solved. Among them, **P** is the desired observer pole vector.

$$\mathbf{G}=place(\mathbf{A'},\mathbf{C'},\mathbf{P})'$$

VI. CONTROL STRATEGY OF MATRIX CONVERTER MOTOR SYSTEM

Control Strategy of Matrix Converter Motor System is essentially a combination of matrix converter control and motor control.

A. Matrix Converter Space Vector and Motor VVVF Control

The current study generally is matrix converter space vector and motor VVVF control. VVVF control obtained the amplitude and frequency of reference

output voltage, matrix converter space vector modulation of matrix converter achieved input power factor correction and output voltage generation. This only on the speed closed-loop, it can not meet the speed requirements of high-performance motor system. Because in the space vector modulation, the duty cycle calculation is entirely based on ideal input and expectations output, it unrelated to the actual input and output and non- suppression to harmonic, robustness of the input, voltage space vector can be introduced as a negative feedback to closed loop control[10].

B. Hysteresis Current Comparative Method

Hysteresis current comparison method can combine with vector control, namely use vector control to the motor and obtain the stator three-phase reference current to control matrix converter by using hysteresis current comparison method, which specific realization can be divided into indirect and direct method by whether to adopt the virtual DC link thought.

Concrete realization of the direct method as follows: When hysteresis current comparator requires to increase a phase current, at this time turn-on the input voltage maximum phase; when hysteresis current comparator requires to decrease a phase current, at this time turn-on the input voltage minimum phase[11].

Indirect method is the matrix converter equivalent to the AC - DC - AC structure of abolition of DC link, in which hysteresis current comparison method control inverter part, and output voltage of the rectifier link Upn is the DC bus voltage of inverter link. DC bus voltage of equivalent AC - DC - AC structure is a three-phase input line voltage and fluctuate range from zero to the maximum of input line voltage. Therefore, the DC bus voltage of equivalent AC - DC - AC structure is present of high, middle and low range of options. The range of high voltage including $[0.866U\max, U\max]$, the range of middle voltage including [0.5Umax, 0.866Umax], the range of low voltage including $[0, 0.5U\max]$, $U\max$ is the maximum of input line voltage.

It can be seen that the direct method is essentially an indirect method of DC bus select the high voltage.

Hysteresis current comparison method set up the premise that the DC bus voltage is greater than the motor stator EMF. When the DC bus voltage of the matrix converter is equivalent to the AC - DC - AC structure is small, there may be turn up the situation of the DC bus voltage less than the stator EMF. At this point, hysteresis current comparison method to actual control on the current effect opposite to expectations and consequently exacerbated, this will increase current and torque ripple and reduce system performance. Therefore, the DC bus voltage chooses generally high voltage.

Simulate research to matrix converter motor system vector control of adopted hysteresis current comparison method using Matlab/Simulink. In which DC bus voltage use high voltage and motor use the three-phase four-pole squirrel-cage induction motor. Rs = 0.087Ω, Rr = 0.228Ω, Ls = Lr = 8mH, Lm = 34.7mH and current hysteresis width is 2A. Initialize torque as the value of 0N.m, 1.5S step to 200N.m; speed set 120rad / s. Speed controller adopt PID controller. Simulation waveform of vector Control of Matrix Converter Motor System is shown in Figure 2 to Figure 3.

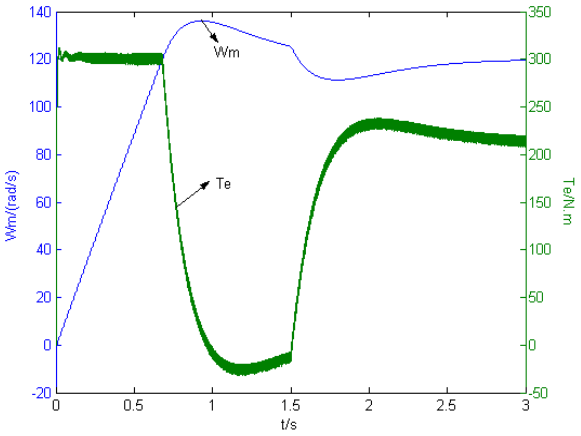


Figure 2. The rev and torque of motor

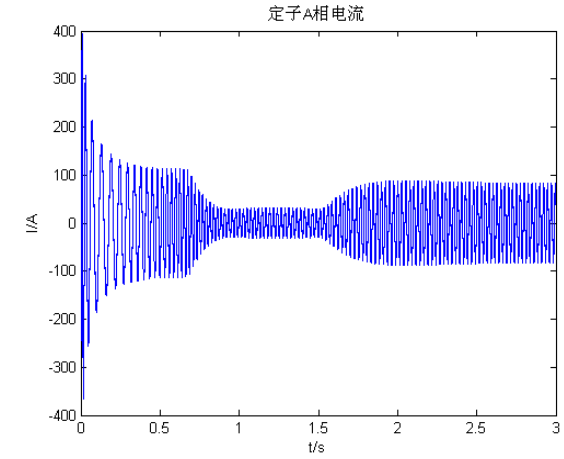


Figure 3. The stator electrical current of A

The control of the rectifier part is only provided DC bus voltage to the inverter, actually we can through adopt space vector modulation to achieved power factor correction to the input the rectifier, but there is a

problem that when the output voltage as middle voltage or zero voltage in the rectification part, hysteresis current comparative method in the inverter failure and thus effect the vector control. The waveform of rectifier link using space vector modulation simulation is shown in Figure 4 to Figure 5.

Space vector modulation can also be combined with vector control, stator reference current obtain by vector control convert into the stator reference voltage, using space vector modulation, input power factor get from rectification links and inverter part generated the requirements reference voltage[12].

C. Direct Torque Control

The combination of direct torque control and matrix converter control can be interpreted using matrix converter output voltage space vector.

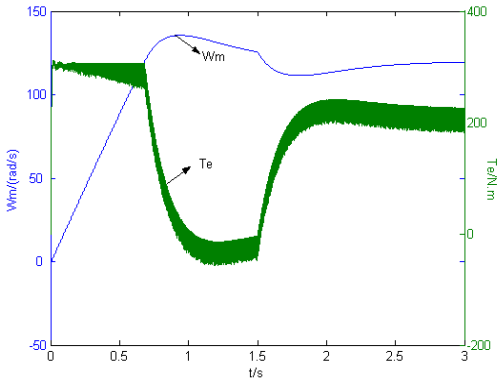


Figure 4. The rev and torque of motor

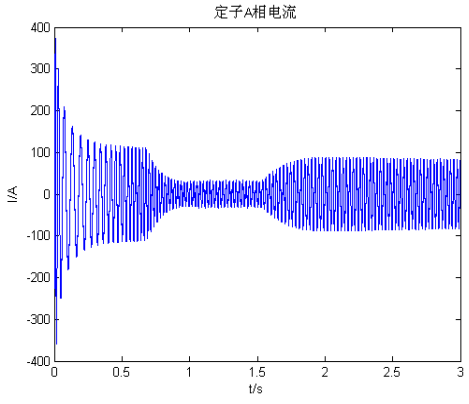


Figure 5. The stator electrical current of A

Three-phase / three phase matrix converter has a total of 27 switch states and corresponds to the 27 voltage vectors. They can be divided into three categories: ①zero voltage vector; ②amplitude changes, voltage vector of constant phase angle; ③ fixed amplitude, voltage vector of phase angle. Direct torque control to qualitative control of torque and stator flux amplitude by using hysteresis comparator. The phase angles of voltage vector determine its nature of the role of change stator flux and torque and amplitude determine the size. The fluctuation of voltage vector amplitude does not affect the feasibility of direct torque control. For the third type of voltage vector, system requires real-time information of voltage vector phase angle to determine its change role to torque and flux. In order to reduce the burden on the system, the matrix converter motor direct torque control system generally select only the first and second voltage vector. Therefore, there are three different voltage vector amplitude has the same phase angle any time, the presence of high, medium and low options. Following paper will analyses high, medium and low voltage vector effect on the system performance.

Because the same sampling time, the higher voltage vector magnitude, the greater torque ripple caused, dynamic performance is proportional to voltage vector amplitude.

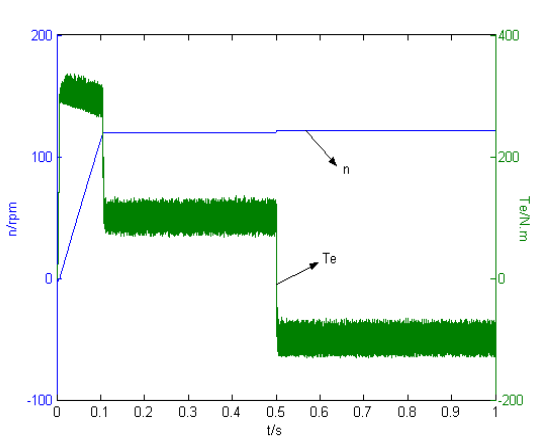


Figure 6. The torque and rev of motor

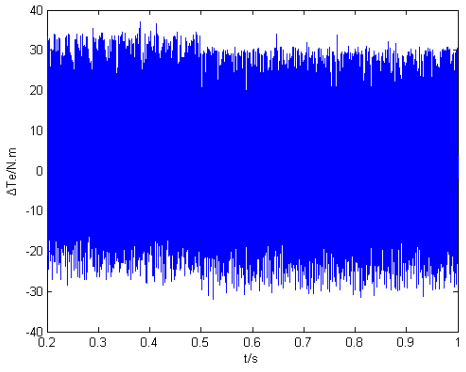


Figure 7. The error of torque ΔTe

The torque ripple of direct torque control system can be divided into two categories: First, in the constant sampling period, the torque ripple caused by discrete is proportional to the input line voltage amplitude. Second, torque ripple caused by the failure of the direct torque control. Direct torque control to select the forward voltage vector and increase rotor flux by increasing the angle between stator and torque. This requires the speed of stator flux rotate faster than the rotor flux. Stator flux rotation speed determined by the input line voltage, so when the input line voltage is small, it will apparent the situation of increases the torque by select a system voltage vector but the actual torque decline contrary. The size of the torque ripple and the frequency is inversely proportional to the input line voltage amplitude. Studies show that the high-voltage vector does not appear the failure situation of direct torque control, while medium and low voltage vectors both appearances. Thus, high-voltage vector generally be chosen. Simulate research to direct torque control by using the matrix converter motor system of high-voltage vector and the simulation results shown in Figure 6 to Figure 7.

By selecting the high and medium voltage vector to reduce the torque ripple of direct torque control system, the specific is that select the input line voltage is high when the system is in dynamic; select the input line voltage is middle when the system is in stable, when in the case of failure of direct torque control, select the input line voltage is high and determine the system state according to the size of the torque ripple. Simulated research to the matrix converter motor direct torque control system using this strategy and the simulation results shown in Figure 8 to Figure 9.

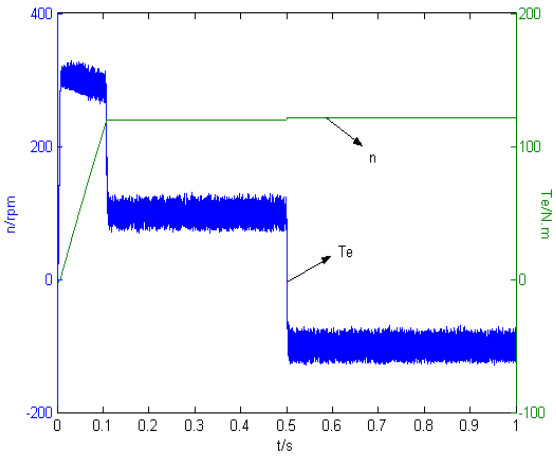


Figure 8. The torque and rev of motor

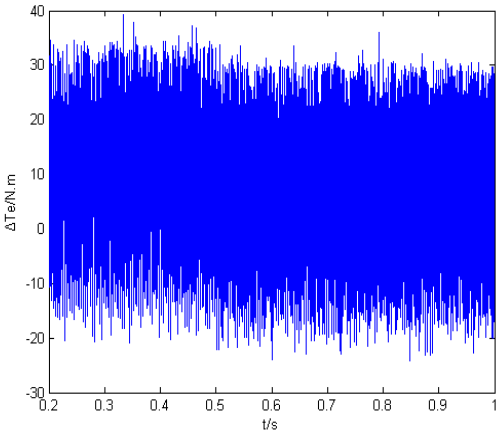


Figure 9. The error of torque ΔTe

By comparison Figure 7 and Figure 9 shows that this improvement strategy can reduce the torque ripple effectively.

In the rectifier link space vector modulation similar to vector control indirect method. The matrix converter is equivalent to the AC - DC - AC structure, control rectifier link using direct torque control and use space vector modulation to inverter link.

VII. CONCLUSION

This paper proposed the concept of matrix converter motor system aimed at the problems of the matrix converter enter into industrial applications and research the matrix converter motor system control strategy combined matrix converter control method and motor speed method, analyzed the impact of the size of the voltage vector to hysteresis current comparator and direct torque control, and proposed to the improved control strategies of reduce the torque ripple and consider the input power factor control.

ACKNOWLEDGMENT

The authors wish to thank the cooperators. This research is partially funded by the Project funds in Shaanxi province University Student Innovation and Entrepreneurship Fund Project (S201910702064) and the Project funds in engineering laboratory project(GSYSJ2018012).

REFERENCES

[1] P. W. Wheeler, R. Jose, C. Jon, et al. Matrix converters: a technology review [J]. IEEE Trans. on Industrial Electronics, 2002, 49(2): 276~288.

[2] Li Yao-hua and so on. an effective way to improve voltage transfer ratio of matrix converter [J]. Electric Automation, 2006, 28 (5): 32 ~ 35.

[3] Guo Qian-gang, Li Yao-hua, Meng Yan-jing etc. the theoretical analysis and discuss of matrix converter voltage transfer ratio [JElectric Automation, 2004, 26 (4): 55 ~ 57.

[4] Huang Ke-yuan, He Yi-kang, Bian Song-jiang. Variable speed, constant frequency, and wind power system research of matrix converter AC excitation [J]. CSEE, 2002, 22 (11): 100 ~ 105.

[5] C. Klumpner, P. Nielsen, I. Boldea, et al. A new matrix converter motor (MCM) for industry applications [J]. IEEE Trans. on Industrial Electronics, 2002, 49(2): 325~335.

[6] A. Alesina, M. Venturini. Analysis and design of optimum amplitude nine-switch direct AC-AC converters [J]. IEEE Trans. on Power Electronics, 1989, 4(1): 101~112.

[7] Wang Yi, Chen Xi-you, Xu Dian-guo. Research to two voltage synthesis matrix converter loop control [J]. CSEE, 2002, 22 (1): 74 ~ 79.

[8] L. Huber, D. Borojevic. Space vector modulated three phase to three phase matrix converter with input power factor correction [J]. IEEE Trans. on Industry Applications, 1995, 31(6): 1234~1245.

[9] Zhang Zhi-xue, Ma Hao. The current control strategy of matrix converter [J]. CSEE, 2004, 24 (8): 61 ~ 66.

[10] Wang Yi, Chen Xi-you, Xu Dian-guo. Research of Space vector modulated matrix converter closed-loop control [J]. CSEE, 2003, 23 (6): 164 ~ 169.

[11] Ge Hong-juan, Mu Xin-hua, Zhang Shao etc. Permanent Magnet Synchronous Motor Vector Control Model and Simulation based on Matrix Converter [J]. Mining Technology University of China, 2005, 34 (3): 333 ~ 337..

[12] Sun Kai, Huang Li-pei, Song Ta Gong-gui. Induction Motor Vector Control based on Matrix Converter [J]. Tsinghua University, 2004, 44 (7): 909 ~ 912.

Review of Network Technology System from the Past, Present to the Future

Mou Chengjin
China International Strategic Research Center of Mobile Communications Joint Association
Beijing, 100029, China,
E-mail: mcjzp139@139.com

Guo Xiaohui
Henan Vocational College of Agriculture
Zhengzhou, 451450, China,
E-mail: mcjzp139@139.com

***Abstract*—Since China's public network got access to the Internet in 1994, the study, research and understanding of the Internet have been blindly superstitious to the United States for a long time, copying the rules and regulations of the United States, and the textbooks in the field of Internet and information are almost completely Americanized. For more than 20 years, we have not formed our own systematic and profound research and practical views on the Internet, most of which are based on the half understood, ignorant, and parrot-like knowledge and cognition instilled from the United States.**

Being controlled by others in ideology is even worse than in technology. At present, China's understanding of the Internet is very superficial from the superior to inferior. We failed to firmly grasp the technology controlled by others and legal key points. We didn't adhere to independent innovation, which resulted in that we were repeatedly passive or even long-term passiveness in cyberspace strategies and tactics.

This paper will start from the history and technical characteristics of the emergence of the Internet, comprehensively discuss the problems that have existed since the emergence of the Internet more than 20 years ago, and reflect on the future development of the international network, making the Internet a truly open and shared international network.

Keywords-Network; Internet; Future Network

I. INTRODUCTION

The computer network refers to the computer system which can provide transmission, storage, analysis and sharing of data for the purpose of acquiring and mastering data. It serves for the needs of communication with others. Two or more computer networks with communication protocol, transmission channel and infrastructure interoperability constitute a computer network interconnection and interworking system that connects and shares data. Strictly speaking, it can be called interconnected network.

Future Network or an international network system consists of all ubiquitous networks connected, interacting and sharing different carriers, sources, function-matching and operating purposes, whether wired or wireless, ground or space. The current Internet is just a computer network using a single TCP/IP communication protocol, not two or more interconnected networks. It is not the interconnection or internet by the exact definition, so it can't be called an interconnected net, or rather it may be called a computer connected system.

If the origin is wrong, everything is wrong. This is especially true of our understanding of interconnected Networks.

What is the Network? What is the interconnected Network? What is the Future Network? What is the different or relationship between IPv4, IPv6 or even IPv9? What are the fundamental drawbacks of Internet architecture and principles? Do we have safe, credible and effective response plans and coordination measures for the newly-emerging problems, new things, new technologies and new spaces in the process of intelligent network development? We all need to advance with the times to re-understand, redefine, re-explore and re-study it. We need to take the opportunity to seriously correct the deviation and mistakes in cognition and knowledge, and let the network users, the people across the country and our future generations know the facts in a practical way.

II. ARCHITECTURE OF THE INTERNET

It is generally believed that the American Internet, which adopts IPv4 technical protocol, has entered the post IPv4 era due to the lack of address design. At present, renting IPv4 address and adopting IPv4 technical protocol constitute the Internet of various

countries, which is still the mainstream of computer network.

The U.S. military says its IPv4 address can be used until the end of 2029. In recent years, the United States has continued to assign IPv4 addresses to the United States and other countries around the world, excluding China.

IPv6 protocol is designed to solve the problem of IPv4 address shortage. It can provide 2^{128} address scale, and that's all. The application and resolution of IPv6 address is still based on the network architecture of IPv4, that is, the original, traditional and irreversible design architecture of the Internet and the tree network architecture (IPv4) continuously improved, strengthened and tightly controlled by the U.S. government and military.

It is inevitable that IPv6 can't interoperate with IPv4 in technology, which is bound to lead to the confusion of network architecture and operation. Therefore, IPv6 special network architecture has to be rebuilt to replace IPv4 network architecture (involving almost all the network software and hardware of infrastructure), which constitutes a "subversion" of the Internet based on IPv4 in fact.

After more than ten years of transitional practice, the United States has found that the cost of rebuilding IPv6 network is too large, there are too many security traps, and the technical agreement is not mature. Besides, "subverting" IPv4 has brought about a series of extremely serious problems in economy, society and military. In fact, the U.S. military and government have abandoned IPv6 transition plans since 2011. In 2017, the adoption rate of IPv6 in the United States dropped from the first in the world to the third.

On July 14, 2017, the Internet Engineering Task Force (IETF) of the United States issued RFC 8200, which announced the latest standard (STD 86) of the sixth edition of Internet Protocol (IPv6). At the same time, it abandoned the RFC 2460 (IPv6 draft) proposed in December 1998, and deleted the "next generation Internet Protocol IPng" which was in transition to IPv6.

Over the past few years, the widespread introduction of new data protection regulations around the world is having a dramatic impact on technology companies and consumers around the world, resulting in some previously established best practices in IETF procedures and regulatory requirements becoming undesirable, the U.S. Internet regional working group (IntArea) said.

Please note that the U.S. Internet Engineering Task Force issued the official document RFC 8179 (BCP 79) in May 2017, namely, "intellectual property issues in IETF technology", which provide three basic principles for IETF to deal with Internet intellectual property claims (discarding RFC 3979 and RFC 4879 simultaneously):

1) The IETF will not determine the validity of any specific intellectual property claim.

2) In following the normal practice? The IETF can decide to use technology that has been exposed as intellectual property if necessary.

3) All participants in the IETF working group discussions must disclose known intellectual property rights or any intellectual property rights covered and likely to be covered under discussion and their recommenders. This requirement applies to all IP claimants, their employers, sponsors, or agents of IP claimants, without the need for patent searches.

In this way, IETF tends to choose technologies with undeclared intellectual property rights, or technologies with free intellectual property rights; IETF may adopt technologies at its discretion, or not commit to licensing; and IETF specifications do not stipulate mandatory security technologies. Therefore, IETF does not define the internal or external problems of the main patent technologies for IPv6.

So what's the practical significance of saying that China has more than 100 IPv6 intellectual property rights? After all we are still subject to the United States and IETF!

III. THE PROBLEM OF IPV6

Practice has proved that many IPv6 Security traps occur and appear when IPv6 cannot interoperate with IPv4, or when IPv6 is trying to run with the network architecture of IPv4 technical protocol. Once they happen, they will not go away, just like opening Pandora's box (security trap or temptation) of network security. For example:

The design of interface ID in IPv6 address will lead to the mandatory real name system for ordinary users in disguise. Because IPv6 also stipulates that interface ID can be allocated in other ways, even random number and manual way, the experts of IPv6 like hackers can easily hide their physical address. This state is no more insecure than IPv4, but an astonishing security scandal once it is widely known, easily operated and arbitrarily adopted by ordinary users who have no knowledge of IPv6. The gateway tunneling

may also help hackers or spies of hostile camps hide their whereabouts, making hackers more difficult to find, or causing greater national strategic security problems.

Network address and addressing mode of IPv6, data routing and exchange are real end-to-end, and there is no need for network address translation (NAT). At the same time, the network identification of user equipment is directly exposed, which can be easily collected and used. Through the cross aggregation and correlation analysis of multi-source and multi-element identification data, it is easy for humans and machines to be bound permanently, and thus beyond the current "precise push" (advertising) ability, deriving "precise tracking", "precise positioning", "precise strike", etc., with great potential security risks. IPv6 is applied to smart home, smart community, big data, cloud computing, etc. It may be "accurate" to the details of a family, a family member or the staff in the same office, etc., which is extremely dangerous.

On the one hand, almost all servers of well-known websites are hosted abroad. For example, Netease e-mail server is hosted on Amazon's cloud service platform (AWS). At least the IP address belongs to Amazon. The risk and consequences of domain name and address being controlled are obvious. It is simply to hand over hundreds of millions of Netease users to the U.S. Central Intelligence Agency and its intelligence system (IC) members for all-round, all-view and all-time monitoring and supervision.

Amazon has publicly announced the provision of cloud services to the CIA and its members of the intelligence system (IC), which is known as the "Amazon cloud service platform secret zone" (AWS, Amazon Web services). Amazon called the service "the first and only commercial cloud provider to provide comprehensive data classification services to the government, including non-secret, sensitive, classified and top secret data."

On the other hand, BIND, the system software of DNS server, has become the standard of implicit monopoly. Almost all users in the world do not know the truth (the relevant national authorities and scientific research institutions have never issued a warning, nor have guided users taken any preventive and governance measures), that is, the United States has long been on all DNS servers (IPv4 and IPv6) on the Internet, solidified the necessary route to the network information center of the United States Department of defense first. No matter what users are who, whether like it or not, all data and information exchanges must unconditionally comply with the security principles and measures of "American interests first".

The Great Wall firewall is invalid for IPv6. At present, IPv6 network in Colleges and Universities can easily log in the "forbidden network" of foreign countries (websites of religion, terror, anti-propaganda, etc.). Another reason why IPv6 is not suitable to replace IPv4 is the conflict about IPv6 on the Internet backbone network, which leads to the congestion of network flow. At present, there is no reliable technical solution.

The overall comparison of IPv4 and IPv6 in the case of a single failure shows that in 86% countries, IPv4 connection is more reliable. An important discovery in IPv6 field is that many ISPs do not have correct network connection under normal operation conditions. For example, in the United States, only about 10% of autonomous systems (AS) support IPv6, while in China, China Telecom (AS 4134) only gets global connectivity through one service provider, hurricane electric (HE), which is in worse condition.

Technically speaking, China's public network has fallen into the hands of others, the above-mentioned major IPv6 Security Risks (pitfalls, temptations and solidified routes, etc.) are not completely solved, and state organs and special departments, as well as important sensitive industries involving the national economy and people's livelihood, dare not use them. If it is used, the consequences will be unpredictable.

The principles, systems, and strategies that US internet is dominated and controlled by the US military remain unchanged. The U.S. military has established and improved a network operation system with strict command and coordination from the top down, especially in the field of cyberspace, which is strictly regulated by the U.S. military.

However, it is difficult for China to make a firm response to network operations in the first time, and to organize a high-speed, high-efficiency and high-intensity anti reaction capability of the military civilian joint network operation system in the first time. The current supervision, command and coordination system and framework of cyberspace in China are neither suitable for the perception situation of the Internet in the United States, which has completed and improved the preparations for launching cyber war at any time in terms of technology and law, nor for the needs of accelerating the construction of cyber power and effectively responding to the United States' overall containment of China in cyberspace.

IV. "TWO CHINA" ON THE INTERNET

ICANN is suspected of deliberately manufacturing "two Chinas" on the Internet for a long time, deliberately setting the technical conditions and basis of "two Chinas" that can cause network information confusion, and deliberately restraining, containing and interfering with China's autonomous and controllable development of sovereign and secure networks.

According to the regulations of internet name and digital address distribution agency in the United States, some IP addresses are assigned to the five regional Internet registries (RIR) in the world, and then the five regional Internet registries are respectively responsible for the registration services in the region. IP address and AS number assignment for Asia Pacific countries are managed by the Asia Pacific Network Information Center (APNIC), which is established in Australia. Under the five regional Internet registries, RIR is divided into national registration agency NIR and regional registration agency LIR. The U.S. Internet name and digital address distribution agency divides the Asia Pacific region into 56 economies (countries and regions).

The Asia Pacific Network Information Center has seven core members (national registration agencies) who can enjoy special preferential conditions, including China, Japan, South Korea, Vietnam, India, Indonesia, and even Taiwan.

According to the official website of Taiwan Internet Information Center (TWNIC), founded in December 1999, its original competent department was the Ministry of Transport of Taiwan Authorities. In December 2017, it was changed into the National Communication and Broadcast Commission and became the national network information center.

In December 1999, it was at a time when Lee Teng Hui publicly supported Chen Shui Bian's campaign for "President" and Taiwan's independence was more active. On May 20, 2000, a gun shot put Chen Shui Bian on the throne of "President". Since then, in the "Internet" activities held around the world, Taiwan's "national flag of the Republic of China" has been put in the venue; Taiwan's representatives of TWNIC, the "national registration agency", enjoy the same treatment as those of CNNIC (China Internet Network Information Center) of the People's Republic of China.

China has clearly declared the principle of "one China" sovereignty in international organizations, but why is the Internet an exception? Why should we tolerate the emergence of "two Chinas" on the Internet many years after China's full-featured access to the Internet in 1994? It concerns the geopolitical issues of Internet governance, the logical, physical and perceptual boundaries of Internet monitoring, and the fact that Taiwan can easily control China's data sovereignty and open-source information by using domain name rotation, data "transgenic" technology and other technologies. It concerns the sovereignty principle and security bottom line of China's cyberspace. This is not a simple technical issue, but a general political issue.

China's "Anti-Segregation Law", issued in 2005, clearly declared that there is only one China in the world, the mainland and Taiwan belong to one China, and China's sovereignty and territorial integrity are inseparable. The state will never allow the "Taiwan independence" secessionist forces to separate Taiwan from China in any name or in any way. The fact that the "Taiwan independence" secessionist forces split China in any name or in any way, the state may take non peaceful means and other necessary measures to safeguard its sovereignty and territorial integrity.

Any negligence on the sovereignty and security of cyberspace data (no matter professional or amateur) may lead to irreparable loss or disaster of cyberspace sovereignty and security (national sovereignty and security) at any time. How can one tolerate others encroaching on one's preserve?

V. SUGGESTION

Firstly, re-understand the Internet and deepen the governance of the Interconnected Network. Based on a wide range of opinions, we should open up and conduct large-scale discussions on the deployment of IPv6, practically adjust our strategies and tactics in the field of cyberspace information, correctly guide and promote the construction and development of China's sovereign network, future network, and the global community of destiny in cyberspace.

Secondly, re-consider the e-government extranet, comprehensive website and network infrastructure security, design and implement China's autonomous and controllable cyberspace security monitoring system.

Thirdly, thoroughly eliminate and eradicate the adverse effects, political weaknesses and technological constraints of "two Chinas" on the Internet.

About the authors

Mou Chengjin, the director of the international strategy research center of China Mobile Communications Federation. The paper version was

firstly published in Chinese on December 12, 2018, revised in January 2020. If the Chinese version was needed, please contact the author. Email: mcjzp139@139.com .

Guo Xiaohui, associate professor of Henan Vocational College of Agriculture. Email: guoxiaohui@hnca.edu.cn

Review of 3D Point Cloud Data Segmentation Methods

Ruan Xiaoyi
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710032, China
E-mail: 2827538117@qq.com

Liu Baolong
School of Computer Science and Engineering
Xi' an Technological University
Xi'an, 710032, China
E-mail: liu.bao.long@hotmail.com

***Abstract*—3D point cloud segmentation is one of the key steps in point cloud processing, which is the technology and process of dividing the point cloud data set into several specific regions with unique properties and proposing interesting targets. It has important applications in medical image processing, industrial inspection, cultural relic's identification and 3D visualization. Despite widespread use, point cloud segmentation still faces many challenges because of uneven sampling density, high redundancy, and lack explicit structure of point cloud data. The main goal of this paper is to analyse the most popular algorithms and methodologies to segment point clouds. To facilitate analysis and summary, according to the principle of segmentation we divide the 3D point cloud segmentation methods into edge-based methods, region-based methods, graph-based methods, model-based methods, and machine learning-based methods. Then analyze and discuss the advantages, disadvantages and application scenarios of these segmentation methods. For some algorithms the results of the segmentation and classification is shown. Finally, we outline the issues that need to be addressed and important future research directions.**

Keywords-Line Point Cloud; Segmentation; Classification

I. INTRODUCTION

Image segmentation is one of the basic research directions of computer vision, and its purpose is to subdivide a digital image into multiple regions with similar properties[1]. Segmentation of 2D images has more than 50 years of research history, but 3D point cloud data is a highly redundant and irregularly ordered structure, point cloud segmentation also faces many challenges. The segmentation of point clouds into foreground and background is a fundamental step in processing 3D point clouds. Given the set of point clouds, the 3D point cloud segmentation can be defined with the sets:

Let the surface set $F = \{F_1, F_2, ...F_n\}$ reconstructed from the point set S, and the set $R = \{R_1, R_2, ...R_n\}$ be a subset of the power set of S. Each element in R is a set of point sets corresponding to a certain characteristic surface in F, which represents a region obtained by segmentation. If the following conditions are met, R is called a partition of the point set S:

1) $\bigcup_{i=1}^{n} R_i = S$, indicates that the union of the divi-ded regions is the measured point set S, that is, each measurement point is divided into a certain region.

2) $R_i \cap R_j = \emptyset$, indicates that the point sets obtained by segmentation do not intersect each other, and each measurement point cannot belong to two different regions at the same time.

3) The points in each region have the same characteristics, and any two adjacent regions do not have the same characteristics.

4) $V_i (i = 1,2,...,n)$, R are all connected regions.

II. METHODS SUMMARY

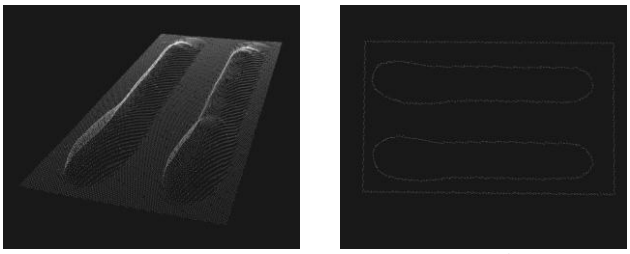
We divide the 3D point cloud segmentation method into:edge-based methods, region-based methods, graph-based methods, model-based methods, and machine learning-based methods based on the basis of the current segmentation.

A. Edge-based methods

The edge-based segmentation method is currently the most studied method[2]. Edges are the basic features that describe the shape of point cloud objects (Figure 1). The edge-based segmentation method first detects the geometric boundary points of the data points. The edge of the point cloud is usually composed of the normal vector of the point cloud or the boundary points with abrupt curvature. These boundary points

are connected and then the entire data set is divided into the independent multiple point sets achieve the purpose of segmenting the point cloud. Therefore, the edge detection technology is the key to edge-based segmentation.

Bhanu[3] first used gradients to detect edges, fitting 3D lines to set points, and detecting changes in the direction of the unit normal vector on the surface. Subsequently, Jiang[4] proposed a fast segmentation method using scanline grouping technology. The scan lines of the distance image are divided into curves and then they are clustered to represent the surface. This method is better than Bhanu 's method in segmentation quality and running time , but it is not suitable for point clouds with uneven density. Sappa[5] proposed a new edge detection method. By performing edge detection on binary maps, extracting closed contours for fast segmentation; Wang et al.[6] proposed a fast point cloud edge detection algorithm. First, the point cloud data is grid-organized to exclude non-edge discrete points. Finally, AlphaShapes judgment conditions are used to fficiently extract edges, which can effectively avoids the problem of extracting outer boundaries and holes.



(a)original point cloud (b) the edge of data a.

Figure 1. Point Cloud and Its Edge

The advantage of edge-based segmentation is that it has good segmentation results for data with strong contrast in different regions, and can detect the edges of different regions very intuitively. The disadvantage is that although it detects all the edges, it is difficult to determine the detection, the relationship between the edge of the target area and the boundary of the target area. In addition, such algorithms are very sensitive to noise and are not suitable for objects with smooth surface changes.

B. Area-based methods

The region-based method classifies nearby points with similar attributes by searching for neighborhood information to achieve the purpose of segmentation. In 1976 Zucker[7] proposed a " Region Grow " should be split for 2D images, the researchers used after being 3D point cloud. The segmentation as shown in follows the Figure 2. Firstly, a small piece of seed region is given to the segmentation target, and the seed region is used as a starting point to search for the neighborhood; the curvature, normal vector, and geometric features of the point cloud are used as standards; if a point meets the criteria for seed growth, the point is incorporated into the seed region, and the process is repeated as a new seed point, until all points are detected, a growth region is formed. The segmentation results are shown in Figure 3. However, this method has the problems of being easily affected by noise and easily causing the problem of segmentation holes and excessive segmentation. [8-9].

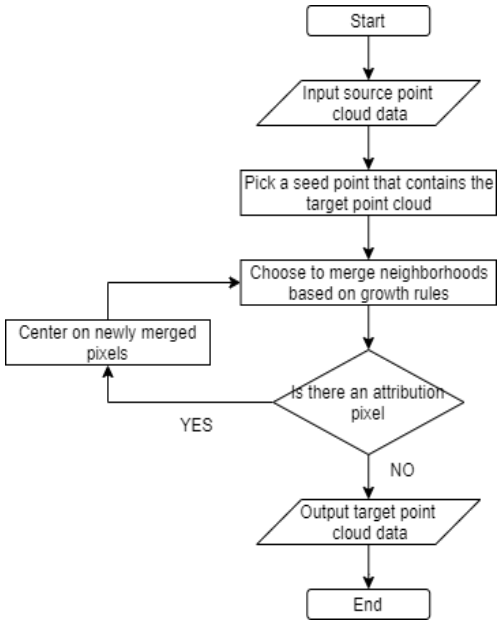
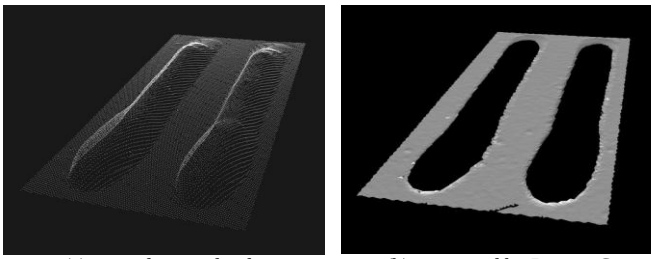


Figure 2. Area-based methods flow chart



(a)original point cloud (b)segmented by Region Grow

Figure 3. Example of Region grow segmentation

Many scholars then the algorithm is improved. Zhang[10] use Otsu determining the optimal segmetation threshold as a constraint condition of growth, better than the traditional segmentation methods, but is susceptible to noise; Angelina et al.[11] improved the region growing method with region merging and genetic algorithms, and its segmentation efficiency has been improved to some extent, but the boundary retention is poor; Xiao et al.[12] proposed clustering-

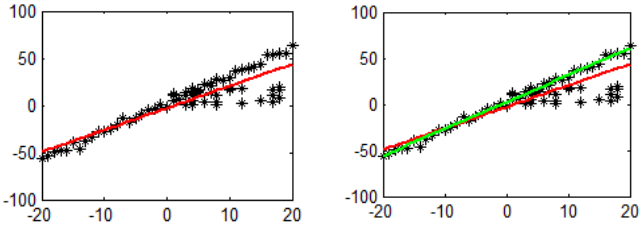
based adaptive region growth for road image segmentation Method, but its scope of application is limited; Besl et al.[13] divide the image into 8 regions based on the curvature characteristics by calculating the Gaussian and average curvature of the point cloud surface, set seed nodes for these 8 regions, and then use polynomial simulation This method is highly susceptible to noise and is time-consuming; Chen et al.[14] calculated the ratio of the minimum eigenvalue of each point and the sum of the three eigenvalues by decomposing the covariance matrix, and the ratio was minimum point as a seed point, the process is mainly used in the building plane extraction rule, the high cost of time; Vosselman et al.[15] use randomly selecting seed sampling points, is determined these seed point whether neighborhood can be fitted to a predetermined model, but the method is prone to false segmentation; Koster[16] using the generated information in a relatively irregular pyramid FIG between the storage area, for comparing and merging adjacent regions; Pu [17] planar surface of the growth surface segmentation algorithm laser transactions; Ning seats[18] proposed a two-step method based on the coarse segmentation and fine segmentation, for extracting the main subject in the scene and finer details.

The region-based method is more accurate than the edge-based method. For the seed method, the segmentation effect is directly related to the selection of the seed points. The quality of the seeds and the merge rules determine the segmentation effect, and this method is extremely susceptible to noise, causes over or under segmentation.

C. Model-based approach

This method is based on geometric shapes such as spheres, cones, planes, and cylinders to group point clouds. According to these shapes, points with the same mathematical model will be divided into the same region. The most typical algorithm is Fischer[19] random sample consensus algorithm (RANSAC) and the improvement of the method. RANSAC achieves segmentation by detecting features with regular geometric shapes such as planes, spheres, and cylinders. The basic principle of the algorithm is to calculate the mathematical model parameters of the data based on a set of sample data sets containing abnormal data to obtain valid sample data. Figure 4 compares the least squares method with the RANSAC algorithm. The results are shown in noisy data In addition, RANSAC can more effectively fit the target.

RANSAC is widely used in the extraction of building surfaces; Li[20] proposed an efficient RANSAC based algorithm on traditional RANSAC, which has good performance. Noise immunity to segment millions of point clouds of a sample in less than a minute. Hough transform the detection points into Hough space[21]. Then a voting algorithm is used to detect objects with a specific shape, so that the problem of detecting arbitrary shapes is transformed into a statistical peak problem[22], which is mostly used for the detection of circles and ellipses. Document[23] compared RANSAC and 3D Hough transform results show 3D Hough transform segment parameter values slower and more sensitive, RANSAC the detection result in the time segments and run more efficiently. However, when RANSAC is applied to plane detection, the "pseudo-plane" problem often occurs. To solve this problem, an improved RANSAC method[24] based on Normal Distribution Transfor-mation (NDT) units, which is accurate. The rate can reach more than 88.5 %, and the plane integrity exceeds 85.0 %. In most cases, the method based on model fitting it needs to provide the geometric model to be detected. Zhang et al.[25] provide a multi-model fitting method based on splitting and merging, which can eliminate the need to set the number of models in advance. In this case, automatic segmentation of point cloud is realized.



(a) Least squares fitting (b) RANSAC

Figure 4. Comparison of Least Squares Fitting and RANSAC

The model-based method has pure mathematical principles, is fast and powerful, and has heterogeneity. The main limitation of this method is the inaccuracy of dealing with different point clouds. This method has been implemented in point cloud libraries based on various models based on lines, planes, circles, and so on.

D. Graph theory-based approach

Graph theory image segmentation uses the principle of graph segmentation to segment the point cloud. This type of algorithm models the point cloud into a graph model composed of nodes and edges that reflect the relationship between the nodes. Graph theory-based optimal segmentation is typified by the FH algorithm proposed by Felzensawalb and Huttenloch[26] . This method constructs a weighted undirected graph using point clouds as vertices, calculates RGB differences

between different vertices to construct weights, and combines regions. Chen[27] achieves the goal of automatic segmentation by building a k-nearest neighbor graph, adding background constraints, and finding the minimum secant line. Compared with several segmentation methods, the results show that this method is suitable for segmenting foreground and background, suitable for specified target extraction, or implementing multi-objective extraction in a supervised classification manner. Ural et al.[28] suppose that a conditional random field model exists, and use graph cut algorithm to disconnect part of the graph model to form independent regions according to the maximum posterior estimation criterion. Li et al.[29] proposed a progressive form of a two-level optimal segmentation algorithm. First, the topological relationship and the distance measurement characteristics of points were calculated under the Riemann geometry frame. The k-means clustering method was used to obtain the segmented voxels as the bottom layer. Segmentation result. Then, the voxels of the point cloud are modeled as nodes, the minimum spanning tree is constructed, and the high-level feature information of the nodes is extracted. The region segmentation effect of the point cloud details is obtained by using graph optimization. Much of the work of graph-based methods has been invested in probabilistic inference models, such as the conditional random field (CRF) method, which uses CRFs to label points with different geometric surface primitives.

Graph theory-based methods can still efficiently segment point clouds in complex backgrounds, but these methods often lack real-time performance[30].

E. Machine learning-based methods

Machine learning treats point cloud segmentation as classifying point cloud data. YU[31] proposed a neural network learning method that combines segmentation and classification simultaneously for target recognition;Yang[32] uses support vector machines to classify geometric features; Guan et al.[33] converts point clouds into voxels data to extract features and then use AdaBoost for training and target extraction; currently the most well-known is Charles et al.[34] use neural network for Point Set for 3D classification and segmentation-PointNet. Although PointNet can discriminate the type of object, it lacks robustness to noise and data.

To solve this problem, Charles et al. Proposed PointNet++, which can extract local features at different scales and obtain deep features through a multilayer network structure. In addition, Zhao et al. [35] proposed a deep neural network model based on multi-scale features and PointNet for feature classification of LiDAR point cloud data in complex scenes. This method improves the local feature rxtraction of PointNet. Ability to realize the automatic classification of LiDAR point clouds in complex scenes; Niu[36] improved the problem of the lack of local topological information of the generated features, and proposed a method that uses bisymmetric functions and spatial transformation networks to obtain more robust and stronger discrimination Compared with PointNet++, the training time is reduced by 20% with the same accuracy.

The advantages of machine learning-based methods are good classification results and high accuracy. But these algorithms require the training process large amounts of data have been tagged to do the training sample, and dividing the object is defined as the target has been trained, it is difficult to point to complete the modeling of the relationship, so that the algorithm is difficult to comprehensively promote.

Comparison of various point cloud segmentation methods is shown in Table Ⅰ.

TABLE I. COMPARISON OF VARIOUS POINT CLOUD SEGMENTATION METHODS

segmentation methods	Advantage	Disadvantage
edge-based methods	Can detect the edges of different areas very intuitively for point cloud.	sensitive to noise and not suitable for objects with smooth surface changes.
region-based methods	More accurate than edge-based methods.	The segmentation result depends on the quality of the seeds and the merging rules. There will be over-segmentation or under-segmentation.
model-based methods	Fast segmentation speed, and heterogeneous,suitable for simple geometric models.	Difficult to use in complex scenarios.
graph-based methods	Suitable for complex scenes.	Lack of real-time.
machine learning-based methods.	Point cloud segmentation has high accuracy, good recognition effect.	lack of real-time.

III. CONCLUSION

Point cloud segmentation is a necessary link for target detection and 3D reconstruction. This paper investigates some classic point cloud segmentation methods and analyzes their advantages and disadvantages. As can be seen from this article, although point cloud segmentation has undergone a long period of research, there are still limitations. Segmentation methods based on edges and regions have good real-time performance but are sensitive to noise, and the segmentation effect is often not ideal. The method based on model fitting has great limitations and it is difficult to adapt to the segmentation requirements of complex environments. Although graph-based and machine learning-based methods can overcome noise and efficiently segment point cloud data, it is difficult to meet real-time system requirements. Therefore, how to overcome the influence of noise and how to improve the real-time performance of the segmentation system is still the focus of point cloud segmentation research.

ACKNOWLEDGMENT

The completion of this paper is inseparable from the patient guidance of Prof. Liu Baolong, Prof. Chen Hua, Teacher Wu Qiong, and Teacher Lipeng Si. I also thank other teachers and students in this laboratory for their help. Finally, I would like to thank the science and technology program of Weiyang District of Xi'an City (Project No.: 2018036) for its fund support.

REFERENCES

[1] Shapiro L, Stockman G. Computer Vision Prentice Hall [J]. 2001:62(3): 271-279.

[2] Luo Xiping, Tian Jie, Zhuge infants, and the like. Summary of image segmentation [D]. 1999.

[3] B. Bhanu, S. Lee, C. Ho, and T. Henderson, Range data processing: Representation of surfaces by edges [J]. In proc.int. Pattern recognition conf. 1896: 236-238.

[4] XY Jiang, H. Bunke, and U. Meier, Fast range image segmentation using high-level segmentation primitives [J], In 3rd IEEE Workshop on Applications of Computer Vision. 1996.

[5] A. Sappa, M. Devy, Fast range image segmentation by an edge detection strategy. In 3D Digital Imaging and Modeling. 2001.

[6] Wang Zongyue, Ma Hongchao, Xu Honggen, et al. Research on water body contour extraction method based on LiDAR point cloud data [J]. Wuhan University Journal of Information Science. 2010:31-26.

[7] Zucker SW. Regiorowing: childhood and adolescence [J], Computer Graphics and Image Processing. 1976, 17(1): 99-125.

[8] Li Renzhong, Liu Yangyang, Yang Man, et al. 3D point cloud segmentation based on improved region growth [J]. Chinese Journal of Optics, 2018:23-26.

[9] Gao F. 375 cases of MATLAB image processing [M]. Beijing: Posts and Telecommunications Press, 2015.

[10] Zhang L, Guo LM, He W, et al. An image segmentation algorithm based on maximal variance between-class and region growing [J] . Information and Electronic Engineering. 2005 .:45-48

[11] Angelina S, Suresh LP, Veni SH K. Image segmentation based on genetic algorithm for region growth and region merging [C]. Computing, Electronics and Electrical Technologies, 2012.

[12] Xiao Xiaoming, Ma Zhi, Cai Zixing, et al. An adaptive region growing algorithm for road segmentation [J]. Control Engineering, 2011, 18(3): 364.

[13] Besl P J, Jain R C. Segmentation through variable-order surface fitting [J]. IEEE Transactions on pattern analysis and machine intelligence, 1988, 10(2): 167-192.

[14] Chen J, Chen BQ. Architectural modeling from sparsely scanned range data [J]. International Journal of Computer Vision, 2008.

[15] Vosselman G, Gorte B G H, Sithole G, et al. Recognising structure in laser scanner point clouds[J]. International archives of photogrammetry, remote sensing and spatial information sciences, 2004, 46(8): 33-38.

[16] Koster K, Spann M. MIR: An approach to robust clustering-application to range image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(5): 430-444.

[17] Pu S, Vosselman G. Automatic extraction of building features from terrestrial laser scanning [J]. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2006, 36(5): 25-27.

[18] X. Ning, X. Zhang, Y. Wang, M. Jaeger, Segmentation of architecture shape information from 3D point cloud[J]. In VRCAI, 2009:77-81

[19] Nguyen A, Le B. 3D point cloud segmentation: A survey[C].2013 6th IEEE conference on robotics, automation and mechatronics (RAM). IEEE, 2013.

[20] Yan Li, Xie Hong, Hu Xiaobin, et al. A new hybrid point cloud plane segmentation method [J]. Journal of Wuhan University (Information Science Edition), 2013:128-130.

[21] Schnabel, Ruwen. Efficient RANSAC for point‐cloud shape detection [J]. Computer graphics forum, 2007, 8(06): 1259-1284.

[22] PVC Hough (1962) Method and Means for Recognizing Complex Patterns, US Patent 3069654.

[23] Tarsha-Kurdi, Fayez, Tania Landes, and Pierre Grussenmeyer. Hough-transform and extended ransac algorithms for automatic detection of 3d building roof planes from lidar data [J], 2007.

[24] Li L, Yang F, Zhu H, et al. An improved RANSAC for 3D point cloud plane segmentation based on normal distribution transformation cells [J]. Remote Sensing, 2017:160-163.

[25] ZHANG Liangpei, Z Yun, C Zhenzhong, et al. Splitting and Merging Based Multi-model Fitting for Point Cloud Segmentation. Acta Geodaetica et Cartographica Sinica, 2018.

[26] Felzenszwalb P F, Huttenlocher D P. Efficient graph-based image segmentation [J]. International journal of computer vision, 2004, 59(2): 167-181.

[27] Chen X, Golovinskiy A, Funkhouser T. A benchmark for 3D mesh segmentation [J]. Acm transactions on graphics (tog), 2009, 28(3): 1-12.

[28] URAL S, SHAN J. Min-cut Based Segmentation of Airborne LiDAR Point Clouds[C]. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXIX-B3. Melbourne, Australia:ISPRS, 2012:167-172.

[29] Li Minglei, Liu Shaochuang, Yang Huan, etc. A Two-level Optimized Lidar Point Cloud Scene Segmentation Method. Journal of Surveying and Mapping, 2018, 47 (2): 269-274.

[30] ANDONI A, INDYK P. Near-optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions[C]//Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science. Berkeley, CA, USA: IEEE, 2006:459-468.

[31] YU Yongtao, LI J, GUAN Haiyan, et al. Learning Hierarchical Features for Automated Extraction of Road Markings from 3-D Mobile LiDAR Point Clouds[J]. IEEE Journal of Selected Topics in

Applied Earth Observations and Remote Sensing, 2015, 8(2): 709-726.

[32] PANG Guan, NEUMANN U. Training-based Object Recognition in Cluttered 3D Point Clouds[C]. Proceedings of 2013 International Conference on 3D Vision-3DV 2013. Seattle, WA, USA: IEEE, 2013:87-94.

[33] Yang Bisheng, Mei Baoyan. Research on 3D City Model Visualization [D]. 2000.

[34] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[35] Zhao Zhongyang, Cheng Yinglei, Shi Xiaosong, et al. LiDAR point cloud feature classification method based on multi-scale features and PointNet [J]. Progress in Laser and Optoelectronics, 2019, 56 (5): 052804.

[36] Niu Chengeng, Liu Yujie, Li Zongmin, et al. Method of 3D target recognition and model segmentation based on point cloud data [J]. Journal of Graphics, 2019, 40 (2): 274-281.