Publisher: State and Provincial Joint Engineering Lab. of Advanced Network Monitoring and Control (ANMC)

Cooperate:

Xi'an Technological University (CHINA)
West Virginia University (USA)
Huddersfield University of UK (UK)
Missouri Western State University (USA)
James Cook University of Australia
National University of Singapore (Singapore)

Approval:

Library of Congress of the United States

Shaanxi provincial Bureau of press, Publication, Radio and Television

Address:

4525 Downs Drive, St. Joseph, MO64507, USA

No. 2 XueFu Road, WeiYang District, Xi’an, 710021, China

Telephone: +1-816-2715618 (USA) +86-29-86173290 (CHINA)

Website: www.ijanmc.org

E-mail: ijanmc@ijanmc.org

xxwlcn@163.com

ISSN: 2470-8038

Print No. (China): 61-94101

Publication Date: September 28, 2022

Associate Editors

Table of Contents

Target Tracking System Based on Dynamic Scene

Fan Yu
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
Xi'an, Shaanxi, China
E-mai: yffshun@163.com

Xue Fan
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district, Xi'an,
Shaanxi, China
E-mai: startfanxue@163.com

***Abstract*—Under the scenario of moving target tracking is very important in the fields of computer vision and research direction, its purpose is to give computer vision, double eye for simulation of the human visual information processing and information feedback process, in the process of the research can obtain the target trajectory information and provide related data for target tracking, The relevant research results of target tracking have been successfully developed in many fields such as human-computer interaction, security monitoring, intelligent monitoring and face recognition. However in the scene moving target vulnerable to disorder interference, light, shade, and scale complex situations, such as the change of the apparent dynamic change in the process of movement target, research can achieve different scenario stable tracking method has important research significance and engineering related events in the field of value, the existing popular nuclear-related filter tracking algorithm was improved, To solve the problem of occlusion and multi-scale change of tracking target, a new tracking algorithm based on multi-feature and multi-scale anti-occlusion is proposed by using kernel correlation filtering method.**

Keywords-Target: Tracking Kernel; Correlation Filter; Fusion Feature; Multi-Scale; Kalman Filter

I. RESEARCH BACKGROUND AND SIGNIFICANCE

Moving object tracking technology has nearly 20 years of research history, the research content of this technology mainly involves moving object detection and extraction, moving object tracking, moving object recognition, moving object behavior analysis and solution and many other aspects, is an important branch of computer vision research. During this period, with the development of computer communication technology, computing technology and information technology, as well as the continuous updating of computer hardware including image processing hardware, moving target tracking technology has also been rapidly developed, and the research results of related technologies have been widely used.

Basic theories such as recognition and artificial intelligence are in turn the focus of these theoretical research. Moving target tracking technology was initially applied in video surveillance system, and then it has been widely developed and applied in practical fields such as national defense, intelligent robot, intelligent traffic command and so on.

Moving target tracking as the digital image processing and computer vision a hot research field in the human production and life of great economic benefit and application value, has been attracted a large number of scholars to research, this paper selected topic in the application background of intelligent video surveillance, focused on analysis of the dynamic tracking targets in the scene change shade and multiscale problems.

The main research content is to realize tracking of moving target in dynamic background. By extracting the histogram of direction gradient and histogram of Lab color space in target information features, the features of the two are fused. Mainly involved in the process of texture feature extraction of target, the target tracking algorithm implementation and tracking the results of the training classifier and is used to determine the target location of several core modules, designed to take advantage of tracking algorithms have been tracking results, and then give a complete target

DOI: 10.2478/ijanmc-2022-0021

mobile state information, determine the goal of tracking, complete experimental analysis.

II. ALGORITHM OF KERNEL CORRELATION FILTER

There are two kinds of target tracking algorithms: generative algorithm and discriminant algorithm. The kernel correlator algorithm belongs to the latter, which is similar to the principle of discriminant algorithm. In its realization principle, classifiers are replaced by filters to judge the position of the tracking target. The positive-negative Sample method is still used in the training of classifier, that is, the training method of Positive and Negative samples. The target region is generally regarded as a positive sample and the area around the target as a negative sample. The closer the location in the image is to the target, the more likely the area around it is to be a positive sample.

A. Characteristics of kernel correlation filter algorithm

1) The circular matrix is cyclically shifted in the area around the tracking target, positive and negative samples are collected, and the ridge regression method is used to train the filter. Since the cyclic matrix can be diagonalized by the Fourier transform matrix, the method of obtaining the cyclic matrix is simplified by the Hadamard product of the vector (the dot product calculation of the elements in the vector), and at the same time the speed and real-time performance of the algorithm is improved by the simplification of the operation.

2) The kernel function can be used to map ridge regression in vector space to non-vector space. If we want to solve the duality problem or conventional constraints in non-vector space, we can also simplify the operation process by using the property that the cyclic matrix can be diagonalized by Fourier transform matrix.

3) Compared with single-channel features (such as grayscale features), kernel correlation filter can adopt features with multiple gradients and histogram channels such as HOG, or color features with three channels such as RGB and HSV, which provides a method for the algorithm in this section to integrate multi-channel features. Two mathematical models are used in kernel correlation filter algorithm.

B. Implementation process of kernel correlation filter algorithm

1) Select the position of the target to be tracked in the first frame of the video sequence, and the selected rectangular frame area will be enlarged by 2.5 times, and the size of the area will be recorded as M * N.

2) Perform cosine weighting operation on the rectangular box of the framed target position, and then calculate the HOG feature operator to obtain the HOG feature graph of n-dimension. Each dimension that makes up the feature is treated as a sample with an input of size m*n, and the sample is denoted x1, x2, x3... XN.

3) Two-dimensional Gaussian function is used to generate label matrix Y, whose size is consistent with sample M *N.

4) Use the nonlinear regression mode (1) mentioned above to calculate k xx, HOG feature is used as input, and a non-linear regressor is trained as shown in (1):

$$\alpha = \frac{y}{k^{xx} + \lambda} \tag{1}$$

In the next frame of the video sequence, an area with the size of M * N is extracted from the position where the target appears in the previous frame as the sample for cosine weighting, and then the HOG feature map and each dimension is denoted as Z1, Z2, Z3, Z4... ZN is taken as the sample input, and the kernel matrix element Kz of the test sample and training sample in the kernel space is obtained by using formula (2). Then the elements together form the kernel matrix Kz, (3), and K xz is obtained, which is the first row of the matrix Kz transpose matrix.

$$k_{ij}^{z} = \phi(z_i)\phi(x_j) \tag{2}$$

$$k^2 = \phi(x)\phi(z)^T \tag{3}$$

5) The following formula (4) is used to obtain the response matrix under the Fourier domain, and then the response matrix F (z) is obtained

through the inverse Fourier transform, as shown in Formula (4) :

$$f(Z) = k^{xz} \odot a \tag{4}$$

6) The position of the maximum response value is found in the response matrix $F(z)$. If the response value of one position in the matrix is greater than the response value of all other positions in the response matrix, then this position is the position of the target in the current frame. However, if the situation is reversed, you can take additional steps (such as re-performing a global search match) to retrieve a target region, and then start again from step (1).

7) The following steps are mainly to update the model, extract samples from the newly found target location, and then repeat the four operations from Step 2 to Step 5. For the current frame, the model is denoted as, then the model used in the next frame can be obtained by interpolation operation between the model of the current frame and the model calculated at the beginning, as shown in Formula (5) :

$$a_{new} = m\alpha_{ola} + (1-m)\alpha' \tag{5}$$

8) HERE, M represents the learning rate, and the value is between 0 and 1, indicating the influence degree of the previous model on the current model. If the value is too large or too small, tracking failure will be caused.

III. MULTI-SCALE ANTI-OCCLUSION TARGET TRACKING ALGORITHM BASED ON KERNEL CORRELATION FILTER

The algorithm designed in this paper is mainly divided into two modules, the tracker target detection and initialization module and the tracker model update module. The detailed realization principle of the algorithm and the intermediate results in the realization process are shown in Figure 1.

The first module is the main process by video each frame and calibration in each frame to track the target track box information to extract the feature information of the target image fusion (HOG + Lab color features) and choose target in a box near the center of the sampling, a model is obtained by training, this model can calculate the response value of pixels in each position the image. At the same time, the kalman filter is initialized when the first frame of the video is played, and the motion information of the target to be tracked is counted from the first frame of the video, so that the location of the target to be obscured can be determined by using the predicted target position information when the occlusion of the tracking target occurs.

The main process of the second module is to judge whether the position of the tracking frame of the previous frame is abnormal, such as out of bounds, when the video sequence is about to play the next frame. If so, adjust the position of the tracking rectangle. At the same time detection originally target tracking box size (scale) is changed, if there is no change, in the previous frame samples near the target, operated by the trained model related to the image, save the response value of each sample point pixel, the highest response values and meet the threshold condition of a given location is the center of the target at this time. Then calculate the total peak value of the target location tracking box through a specific function. Then originally set scale step length, here have the effect of multiple, size of rectangular box and scale values to zoom in and out, get two dimensions, one large and one small calculation of total peak area two scales respectively, then and started to get before the peak size comparison, select out the peak has the largest scale, Finally, the position of the target tracking frame of this frame is adjusted by the scale and the target center position obtained previously, and then the model is trained and updated for relevant operations between the sampling position and the image in the next frame.

A. The steps of HOG feature extraction are as follows

The core of HOG feature is to construct the contour information of the tracking target through histogram on the computationally intensive local range. For image I (x, y), the realization idea of the algorithm in this paper to extract the HOG feature of the tracking target is as follows:

1) In order to reduce the influence of illumination mutation, environmental change and other factors, image information needs to be normalized, such as the normalized Gamma space and color space. Proper compression of images can effectively cope with environmental changes such as illumination and partial area shadow. In addition, in order to improve the real-time performance of the algorithm in this paper, color images are generally transformed into gray images for processing.

2) Calculate the first step degree of the image. The first-order derivative operation can not only weaken the ray mutation interference, but also capture the contour information of the tracking target and the texture information of the image. The gradient direction corresponding to the gradient is calculated as shown in Formula .

3) Perform gradient projection on cells. This reduces sensitivity to changes in the appearance of the tracking target. This step extracts information from the local image. Before obtaining the final feature, the image needs to be divided into several cells, and then the histogram of direction gradient of pixels in each cell is counted and accumulated, so as to obtain the final feature information through mapping.

4) Normalize all cells in the block. Usually the cells in the image are shared over multiple different blocks, and because the normalization operation is performed on different blocks, the results of each operation will be different. The interference factor can be further compressed by normalizing the cell.

5) Collect HOG features. HOG feature is extracted from all overlapped blocks in the image to be recognized, and they are combined into the final feature vector for recognition.

In this paper, PCA-HOG features with dimensionality reduction through principal component analysis (PCA) are used to extract target features. If the resolution of the image in the target region is 100*100, then the target image has a total of 10,000 feature vectors, which is a great burden to the algorithm processing. Therefore, PCA dimensionality reduction method is adopted to minimize the number of feature vectors representing target features and improve the performance of the algorithm.

B. Lab color model

Lab Color model is a color model published by CIE (International Commission on Illumination) in 1976. One of the characteristics of this model is that it has nothing to do with the color display ability of the device image. The appearance of Lab just makes up for the deficiency that RGB and CMYK color models previously appeared must depend on the color characteristics of the device. In addition, compared with RGB and CMYK models, one of the biggest characteristics of Lab model is that the color range it represents is the widest, so RGB and CMYK models can be converted to Lab model.

The Lab color space is shown in Figure 1, with three channels L, A and B in total. Luminance of CHANNEL L changes from green to red in the direction from negative to positive in channel A, and from blue to yellow in the direction from negative to positive in channel B.

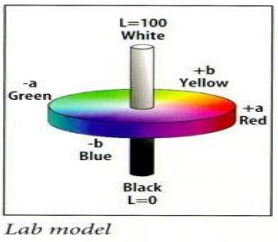


Figure 1. Lab model

Because Lab model has nothing to do with the color display ability of equipment, it has a lot of applications in color image retrieval. In addition, if you want to retain a wide range of color gamut and full color after image processing, the image can be processed using Lab color model, and finally converted into RGB model (for display) or CMYK model (for printing) for output according to demand. This processing method can make the image after processing can try to use more rich and high-quality color output.

To sum up, as a feature, the Lab color model has high sensitivity to color, but low sensitivity to deformation and motion blur, which can be applied to tracking targets when deformation and motion occur.

C. Hanning window and the method of combining the above two characteristics

The kernel correlation filtering algorithm uses Fourier transform, which is often used in time domain and frequency domain conversion. When measuring and calculating digital signals (including images) of infinite length, it is not possible to analyze the whole signal, but rather the limited content extracted from the signal. In the process of signal interception, if the current signal is a one-dimensional sinusoidal signal, if the period of the interception part of the signal is not a positive multiple of the period of the sinusoidal function, then the intercepted signal will be discontinuous on the signal graph, so this interception is called aperiodic interception. It will cause the spectrum of the signal to leak, which is mainly manifested as "trailing" phenomenon in the spectrum. This leakage can be reduced by adding window functions. In order to illustrate the phenomenon of spectrum leakage and the effect of window function to reduce spectrum leakage more vividly, the spectrum diagram of Cameraman images in the digital image standard test set is used to demonstrate.

As shown in Figure 2, the image on the right is the spectrum of the image on the left. As can be seen from the image, there is a thin white line along the vertical axis of the spectrum, which is actually caused by spectrum leakage. As mentioned earlier, when limited interception is performed for digital signal analysis, spectrum leakage will occur if the interception is aperiodic. The mood Conditions are generalized to two-dimensional signals, when using Fourier transform x direction of target sampling sample cycle continuation and continuation, y direction cycle from the point of figure 3, vehicle traffic on the roof of the sky is bright, but driving wheels of the road is darker, the corresponding spectrum produced gray mutation in the y direction, The white line in the y direction appears on the spectrum. Observing the longitudinal edge, the gray level change is not very large, so the white line in the X direction is not very obvious, which also indicates that the spectrum leakage in the X direction is not very obvious. But in general, the presence of white lines represents a leak in the image spectrum.

Figure 2. Car2 image and other spectra

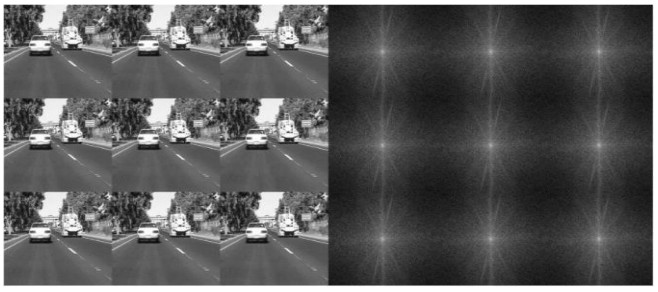


Figure 3. Results and spectra of transverse and longitudinal cyclic assignments of Fourier transforms

To solve this problem, this demo uses a cosine window (a window function) to solve the problem of spectrum leakage. The effect of the image with the cosine window is shown in Figure 4, and the change of spectrum is shown in Figure 5. After the cosine window is added to the image, the phenomenon of spectrum leakage along the Y-axis is not so obvious (the white line becomes lighter), and the energy in the spectrum diagram is more concentrated and clear than before. However, it should be emphasized here that the window function can only reduce the spectrum leakage and will not make the spectrum leakage disappear.

Figure 4. Display effect of Car2 image with cosine window

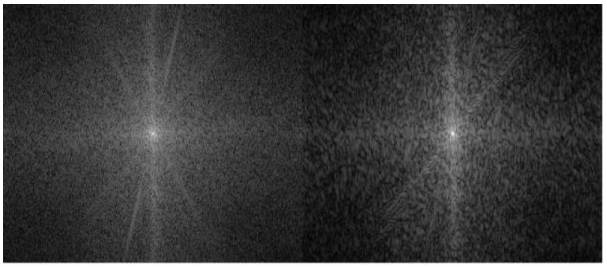


Figure 5. Image spectrum changes after Car2 image plus window

To sum up, the function of cosine window is to enhance the judgment of sample collection in the training model of KCF algorithm, and avoid the phenomenon of training wrong models due to the inaccurate samples collected by cyclic shift.

D. Multiscale introduction and application in this algorithm

Scale space theory: observing the form of any object in daily life, there is a certain measurement standard, this measurement standard is called scale. The measurement of the scale of an object can be understood from many aspects. The weight of an object can be described in terms of "kilogram", "ton" and "kilogram", and the size of an object can be described in terms of "meter", "kilometer", "millimeter" or even "nanometer". When analyzing the unknown scene with machine vision technology, the computer does not know the appropriate scale of the target to be tracked in the image in advance. Therefore, it is necessary to consider the use of appropriate multi-scale image description method (that is, image description at all scales) to obtain the best scale of the object of interest. The invariance of scale should be maintained while obtaining the optimal scale, which means that at different scales, the tracked targets all have the same characteristic key points, so the same key points can be detected for matching for image input of different scales.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental data set and development environment Settings

In the experiment, the OTB-100 dataset was used, which is a set of image sequences specifically used for target tracking algorithms. The entire dataset consists of 100 folders. Each folder contains a video frame sequence set and a tracking target actual position reference text groundtruth_rect. TXT which records the location of the target actual tracking frame when the video frame sequence is played.

In the experiment of the algorithm in this paper, 46 video frame sequences are extracted from the data set to test the algorithm in this paper. These video frame sequences contain various challenges in target tracking, such as scale change, target occlusion, background interference and so on. The programming language of the algorithm in this paper is C++, which is implemented on Visual Studio 2017 software integrating Opencv3.4.1 and opencv-contrib development libraries. At first, the algorithm needs to read the actual tracking frame position determined by the target in the first frame from the reference document of the target position in the video sequence, and then depends on the algorithm to determine the tracking frame position of the target in the video frame. The threshold of occlusion judgment was initially set as 0.25, and the experimental platform was windows10 operating system with Intel i5-6200u 2.40ghz processor. It has 8GB of ram and 2GB of separate video memory. A C++ program generated by static Release compilation can run at a maximum frame rate of 94fps.

The above 46 video frame sequences were used to test the algorithm in this paper and the tracking algorithm of the same type as the algorithm in this paper. The tracking effect and performance curve generated by the algorithm were qualitatively and quantitatively analyzed by Visual Tracker Benchmark. It is realized by matlab programming language. In order to integrate the algorithm into the benchmark tool, the executable file of the algorithm is generated by using the static compilation mode of Release under Visual Studio tools, and then copied to the specified folder. Use MATLAB language to write call interface file, using DOS command to call the executable file, other algorithm implementation program files are also added to the specified folder. Set the video sequence and tracking algorithm to be traced in the configuration file. After that, the rectangle-frame tracking and performance curve drawing functions are executed to describe the rectangle-frame tracking effects of different algorithms and

the tracking performance curves of different algorithms.

B. Qualitative comparison with other tracking algorithms

Due to the large number of video sequences used for testing, there are 46 sequences in total, so four of the most challenging tracking sequences are selected, namely singer1, Singer2, blurFace and Freeman4, which respectively contain deformation, lighting change, blur, fast movement and chaotic background. In the case of out-of-plane rotation and other tracking challenges, the tracking effects of selected and compared tracking algorithms on these 10 sequences are analyzed and compared. At the same time, six filtering tracking algorithms are selected and compared with MulitpleKCF algorithm in this paper, which are CXT, KCF, Struck, CSK, DSST and STC algorithms.

TABLE I. CHALLENGES IN DIFFERENT VIDEO SEQUENCES

Challenge / video	deformation	Illumination	Motion blur	Fast moving	Background clutter	Out of plane	In-plane rotation
freeman4	x	x	x	x	x	√	√
singer1	x	√	x	x	x	√	x
singer2	√	√	x	x	√	√	√
blurFace	x	x	√	√	x	x	√

As shown in Table 1, different tracking challenges in 4 video sequences are presented. If there are corresponding challenge attributes, they are marked as √, and if there are not, they are marked as ×. The tracking effects drawn by this algorithm and other algorithms using Visual Tracker Benchmark tool in these 5 video sequences are shown and analyzed below

Figure 6. Tracking effect of target in Freeman4 sequence

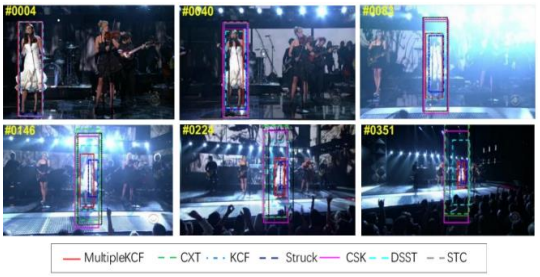


Figure 7. Tracking effect of target in singer1 sequence

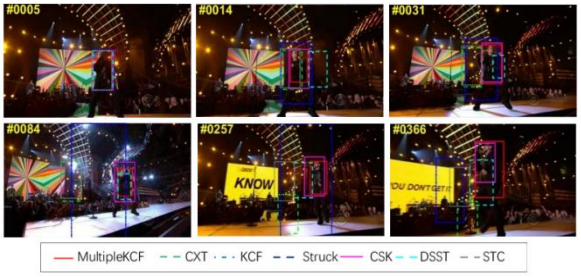


Figure 8. Tracking effect of target in singer2 sequence

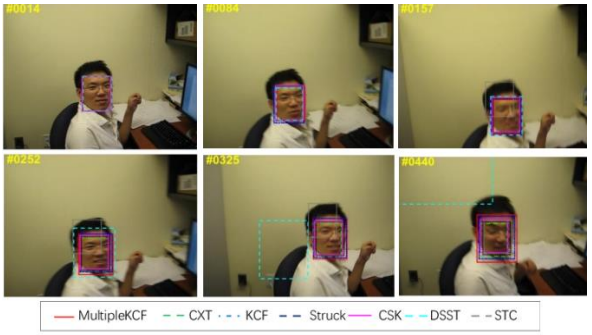


Figure 9. Target tracking effect in the blurFace video sequence

C. The accuracy and success rate of the tracking algorithm are evaluated

Accuracy chart: The center position error is one of the evaluation criteria to measure the accuracy of the tracking algorithm. It refers to the Euclidean distance between the real position of the target calibrated manually and the center position of the rectangular frame obtained by the target tracking algorithm. For each frame used to track the video sequence, assume that the center coordinate of the position rectangular frame calibrated manually is (x,y), and the center target of the target position rectangular frame determined by the tracker is (z, W), for this frame

The center position of is calculated by formula (6)

$$\varepsilon = \sqrt{(x-y)^2 + (z-w)^2} \tag{6}$$

Success rate chart: Another algorithm standard to measure the accuracy of the algorithm is the boundary box overlap rate. The calculation formula of the boundary box overlap rate S is shown in (7):

$$S = \frac{|R_G \cap R_T|}{|R_G \cup R_T|} \tag{7}$$

OPE a traditional tracking algorithm evaluation method, is adopted in this paper. The accuracy and accuracy curves are drawn for the rectangular box position results of 46 video sequences selected by different algorithms after tracking, and the overall accuracy maps and accuracy maps of different tracking algorithms are obtained.

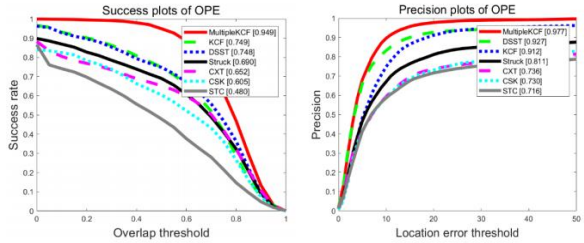


Figure 10. Shows the success rate and accuracy of all video sequences evaluated using OPE

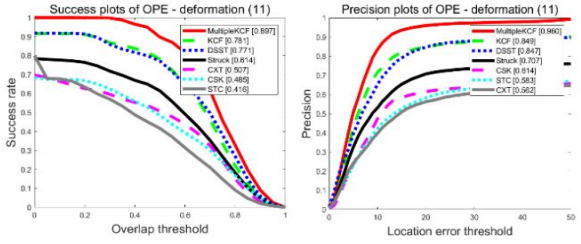


Figure 11. Success rate and accuracy of OPE evaluation under target deformation

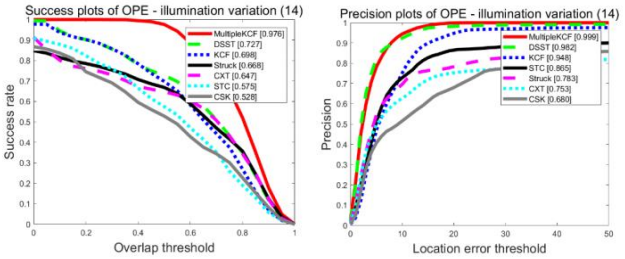


Figure 12. Success rate and accuracy of OPE evaluation under varying lighting conditions

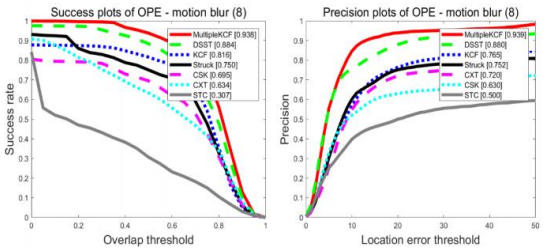


Figure 13. Success rate and accuracy of OPE evaluation in motion blur case

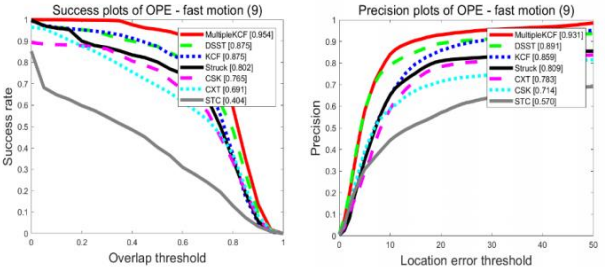


Figure 14. Graph of success rate and accuracy evaluated using OPE in the case of fast movement

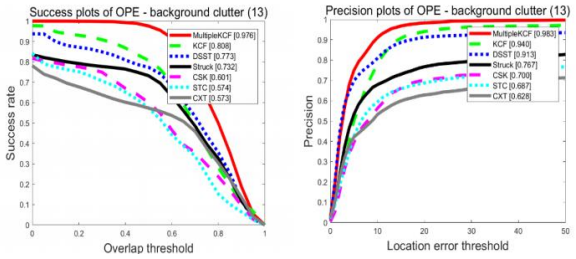


Figure 15. Success rate and accuracy of OPE evaluation in cluttered background

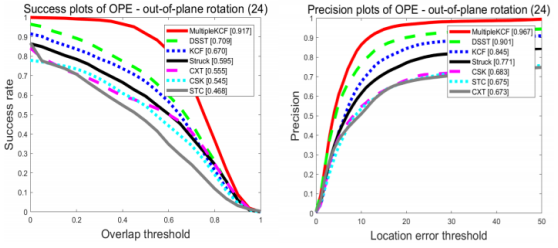


Figure 16. Success rate and accuracy of OPE evaluation for out-of-plane rotation

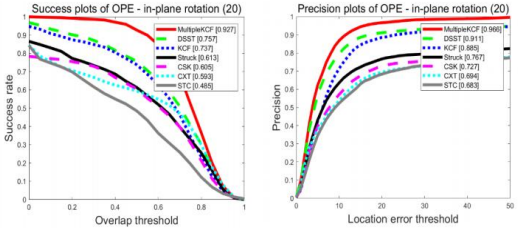


Figure 17. Success rate and accuracy of OPE evaluation for in-plane rotation

By observing and comparing the above curves of performance of different algorithms under different tracking challenges. Performance curves can be analyzed from two perspectives: First, whether it is the success rate or accuracy curve graph of the overall video sequence or individual challenges, it can be seen from the meaning of the horizontal and vertical coordinate parameters and the performance curve legend corresponding to the tracking algorithm: As the abscissa threshold changes, the curve positions of the proposed algorithm are all above those of other filtering algorithms, and its tracking accuracy and success rate are better than those of other algorithms.

In addition, for each success rate graph or curve, in the upper right corner, in addition to the legend corresponding to the algorithm to represent the ranking, there are also their corresponding success rate value or accuracy value, the calculation method has been mentioned in the beginning of this section, so it will not be repeated here. According to the statistics of all the success charts, in different tracking challenges, the success rate of the first ranking tracking algorithm (the algorithm in this paper) is generally 1%-30% higher than that of the second ranking tracking algorithm. In some tracking challenges, such as in-plane rotation, occlusion, Scale variation even reaches 34.2%, 42%, and 62.4% respectively; For all the accuracy graphs, the tracking algorithm ranked first was generally 1%-20% more accurate than the tracking algorithm ranked second under different tracking challenges.

V. SUMMARY

From the above two points of view to analyze the curves of various challenges, it can be concluded that the algorithm proposed in this paper is better than other filtering algorithms compared with it in the success rate and accuracy of two performance indicators.

REFERENCE

[1] Wang H, Wang H, Zhu X, et al. Three-Dimensional Reconstruction of Dilute Bubbly Flow Field With Light-Field Images Based on Deep Learning Method [J]. IEEE Sensors Journal, 2021, PP(99):1-1.

[2] Tokanai K, Kamei Y, Minokawa T. An easy and rapid staining method for confocal microscopic observation and reconstruction of three-dimensional images of echinoderm larvae and juveniles [J]. Development, Growth & Differentiation, 2021, 63.

[3] Kim J, Lee D, Doh G, et al. Three-dimensional tomographically reconstructed optical emission profiles of Hall thruster plasmas [J]. Plasma Sources Science and Technology, 2022, 31(1):015013 (11pp).

[4] Donati D M, Frisoni T. Implant Reconstruction of the Pelvis: IV: 3D-Printed Custom-Made Prosthesis. 2022.

AMM, BAFA, CMJGA, et al. Intraoperative three-dimensional bioprinting: A transformative technology for burn wound reconstruction [J]. Burns, 2022.

[5] AHK, ACL, ASK, et al. Three-dimensional volume reconstruction from multi-slice data using a shape transformation. 2022.

[6] XuY, NanL, ZhouL, et al. HRBF-Fusion: Accurate 3D reconstruction from RGB-D data using on-the-fly implicits[J]. 2022.

[7] Liu Y, Zhang S. Mediastinal basal pulmonary artery identification and classification by three-dimensional reconstruction [J]. Surgical and Radiologic Anatomy, 2022, 44(3):447-453.

[8] GuillouxYL.THREE-DIMENSIONAL RECONSTRUCTION METHOD USING A PLENOPTIC CAMERA:, US20190156501A1[P]. 2019.

[9] K Zieliński, Staszak R, Nowaczyk M, et al. 3D Dense Mapping with the Graph of Keyframe-Based and View-Dependent Local Maps [J]. Journal of Intelligent & Robotic Systems, 2021, 103(2).

[10] Cong M, Fedkiw R, Lan L. OBTAINING HIGH RESOLUTION AND DENSE RECONSTRUCTION OF FACE FROM SPARSE FACIAL MARKERS:, US20210150810A1 [P]. 2021.

[11] Weilharter R, F Fraundorfer. HighRes-MVSNet: A Fast Multi-View Stereo Network for Dense 3D Reconstruction from High-Resolution Images [J]. IEEE Access, 2021, PP(99):1-1.

[12] Lombardi M, Savardi M, Signoroni A. DenseMatch: a dataset for real-time 3D reconstruction [J]. 2021.

[13] Pan Z, Hou J, Yu L. Optimization algorithm for high precision RGB-D dense point cloud 3D reconstruction in indoor unbounded extension area[J]. Measurement Science and Technology, 2022, 33(5):055402 (15pp).

[14] Wang P, Shi L, Chen B, et al. Pursuing 3D Scene Structures with Optical Satellite Images from Affine Reconstruction to Euclidean Reconstruction[J]. 2022.

[15] ShakeriM, Loo Y, Zhang H. Polarimetric Monocular Dense Mapping Using Relative Deep Depth Prior [J]. 2021.

[16] Hller B, Mossel A, Kaufmann H. Automatic object annotation in streamed and remotely explored large 3D reconstructions [J]. Computational Visual Media, 2021.

[17] Theu L T, Tran Q H, Solanki V K, et al. Influence of the multi-resolution technique on tomographic reconstruction in ultrasound tomography[J]. International Journal of Parallel Emergent and Distributed Systems, 2021.

[18] Hermann M, Ruf B, Weinmann M. REAL-TIME DENSE 3D RECONSTRUCTION FROM MONOCULAR VIDEO DATA CAPTURED BY LOW-COST UAVS [J]. 2021.

[19] Farsangi S, Naiel M A, Lamm M, et al. Rectification Based Single-Shot Structured Light for Accurate and Dense 3D Reconstruction [J]. Journal of Computational Vision and Imaging Systems, 2021, 6(1):1-3.

A Path-Wise Scheme for Simpler Mesh-of-Tree Model in Network-on-Chip Designs

Yu Lu
College of Science and Engineering
James Cook University
Cairns, QLD 4870, Australia
E-mail: yulu@gmail.com

***Abstract*—The needs of the chip are growing on a daily basis as extremely large size integrated circuits evolve fast. As a result, the network-on-chip (NoC) was suggested and successfully developed to address the issue of communication on-chip. Current research focuses on NoC topology and routing algorithms that have a substantial influence on network performance. Because low power is a priority, the Mesh-of-Tree (MoT) architecture is preferable. However, certain MoT routers may perform extensive communication activities, which means they consume more power and are more likely to become hotspots. To improve power consumption, we presented a Simpler Mesh-of-Tree (SMoT) architecture based on MoT in this study. We also suggested a path-wise routing scheme for the Internet of Things, which takes into account both network traffic and the shortest path between two points. By altering the direction of data forwarding to reduce network latency, our methodology spreads communication over the whole network. When compared to MoT, our suggested strategy can lower network power consumption by 5.39% -23.3% while also reducing network latency by 4.23%–27.28%.**

Keywords-Noc; Low Power; Routing Scheme

I. INTRODUCTION

Because of its increased efficiency and the ongoing expansion in the number of computational and storage blocks integrated in a single chip, NoC has become the trend of high-performance microprocessors. Many studies are now focusing on NoC design, including topology and routing strategy. Mesh, cmesh, torus, and MoT [1] are some of the topologies that have been proposed. Mesh has gotten the most attention because of its lower node degree. However, due of its larger diameter, it has a higher power consumption.

Root, stem, and leaf routers are used in the MoT architecture, which has a smaller diameter and lower node degree. Root routers may perform intensive communication activities on the Internet of Things, which means they use more power and are more likely to become hotspots. We proposed the SMoT topology in this study, which would diminish this phenomenon while reducing power usage. We also suggested a path-wise routing method for SMoT, which spreads communication around the network by altering packet forwarding direction to reduce network latency.

II. RELATED WORK

Several recent studies have aimed to create a network with low latency and low power consumption. Node-Router Decoupling was suggested by Chen et al., who used power-gating techniques to reduce power consumption [2]. Ghosal, Prasun, and colleagues proposed a Level-2 Mesh topology with two tiers of linkages, one for long-distance communication and the other for short-distance communication [3]. When compared to 3D FMT, Viswanathan designed a revolutionary 3D structure with a 75 percent reduction in the number of vertical links [4].

To minimize network latency, Koibuchi et al. created random shortcuts by supplementing conventional topologies with random links [5]. By delivering packets through less crowded routers, Ebrahimi et al. presented an adaptive routing algorithm [6]. To decrease power dissipation, Ezz-Eldin et al. introduced a low leakage power switch using power supply gating and adaptive virtual channel technique [7].

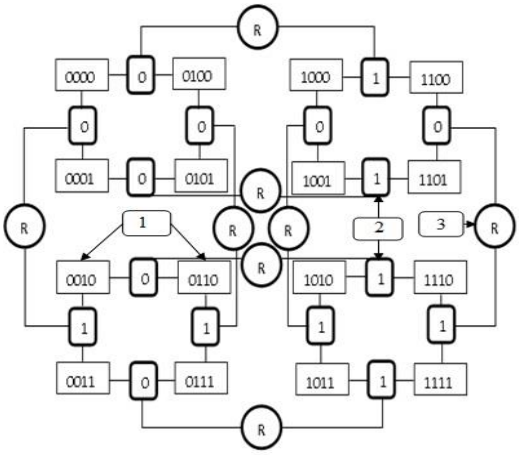
III. MOTIVATION

One of the most fundamental restrictions in NoC designs is power consumption. The linked network of an MIT Raw processor with 16 PEs consumes 36 percent of the total power consumption. The Tera-scale chip was created by Intel, and its associated power usage accounted for 40% of the total power consumption. As a result, NoC has a high-power efficiency. Meanwhile, as process technology advances, communication between cores will become a constraint in further improving performance. Take the Intel Single-Chip Cloud Computer for example (SCC). Each SCC chip has 48 Pentium processors that were coupled by a 46 mesh. It takes at most 10 hops to send a packet from the source to its target node.

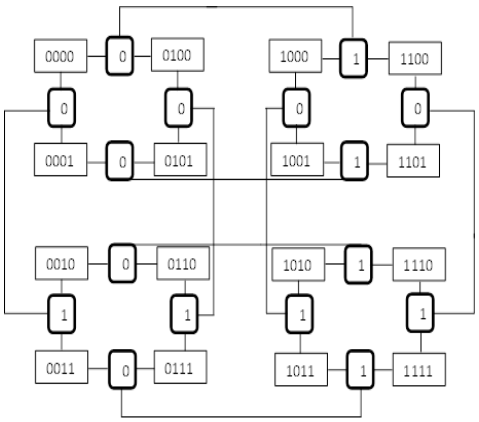
As a result, we presented a more power efficient SMoT architecture in this study. We also proposed a path-wise routing method for SMoT to address the requirement for reduced network latency. Our approach allowed us to reduce the cost of communication, including electricity and latency.

IV. ARCHITECTURE OF THE SIMPLER MESH-OF-TREE

There are three types of routers in the MoT topology: root, stem, and leaf. Row or column routers are the stem routers. Two cores, one row router, and one column router are connected by each leaf router. Until the final level stem routers join the root routers, the stem routers are structured in a binarytree-like design.



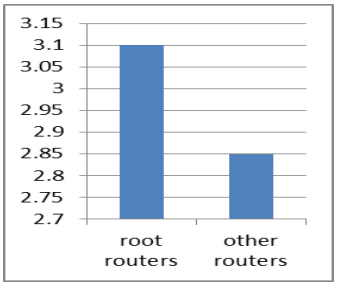
(a)



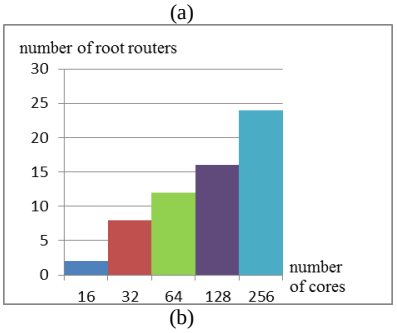
(b)

Figure 1. (a): 32-core MoT (1:leaf, 2:stem, 3:root); (b): 32-core SMoT

Based on the original design depicted in Figure 1(a), we proposed the SMoT, which uses connections instead of root routers to minimize power consumption and latency. For starters, when a network's traffic increases, root routers can quickly become hotspots. The average link use of root and other routers is shown in Figure 2(a). Because other routers have more ports than root routers, packets that arrive at a particular router might pick an appropriate port to be received or sent, or just wait for a while in the buffers. Furthermore, if a significant number of cores are merged, the number of root routers will increase. Figure 2(b) depicts the root routers' growth as the number of cores grows.



(a)



(b)

Figure 2. (a) The average link use of root and other routers; (b) Depicts the root routers' growth

V. PATH-WISE SCHEME

For SMoT, we describe a path-wise routing strategy. When a packet arrives at a leaf router, calculate the congestion coefficient for each neighboring router, which shows the traffic surrounding this router, and select the next hop from the shortest path with the lowest congestion coefficient. This method can minimize network congestion while also providing the shortest path.

In our approach, different routing strategies are applied to the leaf and stem routers, and a few fundamental parameters are required. The number of row routers that a packet passes through on its way from the current node to the destination is represented by Δx, while the number of column routers that a packet passes through on its way from the current node to the destination is represented by Δy. In a M*N MoT structure (M and N denote the number of rows and columns respectively), we also employ a simpler addressing method.

Core_id	Col_ADDR	Row_ADDR
$\log_2 C$	$\log_2 N$	$\log_2 M$

(a)

Col_ADDR	Row_ADDR
$\log_2 N$	$\log_2 M$

(b)

Figure 3. (a) The address for a core; (b) The address for a leaf router

The following is the procedure of the proposed algorithm for leaf routers:

Step1: Packet input.

Step2: Compute the optimal port for packet output.

Step2.1: Calculate the value of Δx and Δy:

$$\Delta x = Des\_addr_{row\_addr} - Src\_addr_{row\_addr} \quad (1)$$

$$\Delta y = Des\_addr_{col\_addr} - Src\_addr_{col\_addr} \quad (2)$$

Step2.2: If Δx≠0 and Δy≠0, which indicates the source and destination nodes are in separate rows and columns, choose one neighbor router as the output:

Step2.2.1: Compute the feasibility P1 and P2 that a packet is sent to the router in the next row or column:

$$P_1 = \frac{\Delta x + P_{algorithm}}{\Delta x + \Delta y + 2P_{algorithm}} \quad (3)$$

$$P_2 = \frac{\Delta y + P_{algorithm}}{\Delta x + \Delta y + 2P_{algorithm}} \quad (4)$$

Where Palgorithm denotes the total number of pathways between a packet's source and destination node, in our proposed SmoT model, Palgorithm is set to 2.

Step2.2.2: Calculate the congestion coefficient for neighbor routers other than the one where the packet was sent, using the equation below:

$$Con_i = P_i \times \frac{Buf\_SIZE_{out1} + Buf\_SIZE_{out2}}{2} \quad (5)$$

Where Coni is the produced congestion coefficient, Pi denotes the probability that a packet is sent to the row router or column router, i=1, 2; Buf_SIZEout1 and Buf_SIZEout2 reflect the number of available buffers in the current output port as well as the number of available buffers in the next available output port.

Step2.2.3: Choose a router with the optimal Coni as the output stop for the packet, along with its output port.

Step2.3: If Δx=0 and Δy≠0, which indicates the source and destination nodes are in identical rows, identify the relevant output port for the row router as the next stop.

Step2.4: If Δx≠0 and Δy=0, which indicates the source and destination nodes are in identical columns, identify the relevant output port for the row router as the next stop.

Step2.5: If Δx=0 and Δy=0, which indicates the source and destination nodes belong to a single router, if Core_id=0, forward the incoming packet to L_Core; if Core_id=1, forward the incoming packet to R_Core.

Step2.6: When two or more packets fight for the same output port, the packet with the larger number of hops wins.

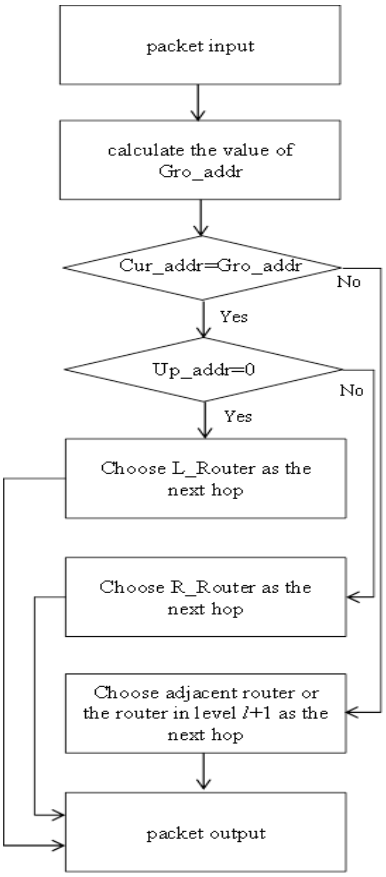


Figure 4. The flow diagram of routing algorithm for leaf routers

Step3: Forward the packet via the determined output port.

Meanwhile, the algorithm for l-th stem routers is presented in Figure 4. Specifically, when a packet comes in, compute the target router in level l. if the present router's location is not the same as the address of the target router, select (l+1)-th router or l-th neighbor one as the next stop, with the condition of current router's level. If not, calculate the address of the target router in level l-1, forward the packet to L_Router or R_Router based on the target router.

VI. EXPERIMENT AND EVALUATION

A. Experimental setup

In this section, we choose two metrics to evaluate the performance of SMoT, i.e., the power consumption and network latency. These two metrics are well adopted in previous works.

In the evaluation of performance, power consumption is a significant factor. It will be meaningful if we can lower it while obtain comparable performance. Overall, the power comsumption is computed as follows:

$$Network\_power = P_{dynamic} + P_{static}$$

The time it takes for a packet to travel from its source to its destination is known as network latency in NoC. One of the primary performance concerns we utilize here is the average latency.

$$Network\_latency = Total\_Cycles / Total\_Flits$$

B. The experimental environment

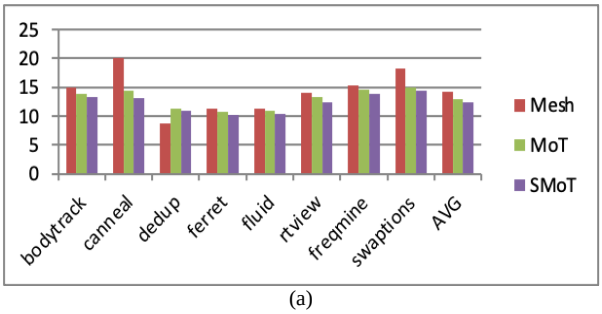
We conduct experiments with the Gem5 simulator, which is a public well-known platform. We choose the PARSEC dataset for evaluation following the the common practice in prior stuides. To be concrete, the details of experimental setup is depicted in Table 1.

TABLE I. ENVIRONMENTAL SETUP

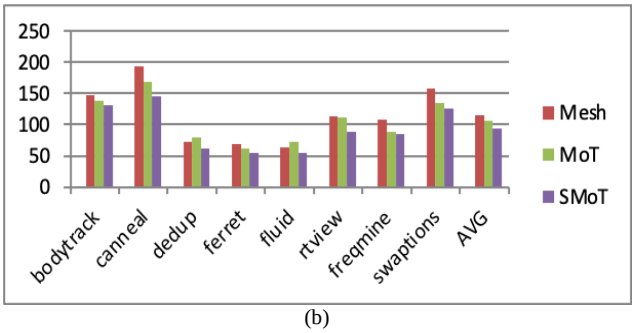
Processor configuration			
CPU_TYPE	Timing	Timing	
CPU_NUMS	16	64	
NoC configuration			
TOPOLOGY	Mesh	MoT	SMoT
SCHEDULING_POLICY	xy routing scheme	deterministic routing scheme	path-wise routing scheme
BENCHMARK	PARSEC	PARSEC	PARSEC

C. Experimental results

The network power comparison is presented in Figures 5(a) and 5(b). As is observed, our proposed SMoT has apparently gained lower power consumption. Specifically, it reduces the power consumption by 5.39%-23.3% compared to the original MoT. We argue this is mainly due to two fundamental factors. On one hand, leveraging short paths between the subnets is able to decrease the number of hotspots, since most routers hold more ports compared to the root ones, packets would obtain more choices for paths. On the other hand, the power consumption is directly reduced with less routers.



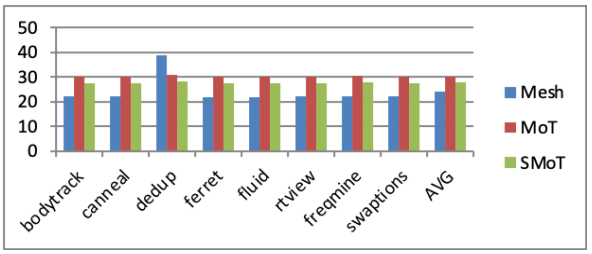
(a)



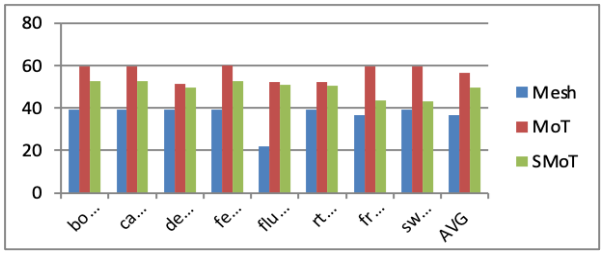
(b)

Figure 5. (a): Network power for 16 cores; (b): Network power for 64 cores

Meanwhile, the results of network latency is shown in Figures 6(a) and 6(b). It is noticed that the network latency of the proposed SMoT was decreased by 4.23%–27.28% than a typical MoT. We believe the reason is that the path-wise scheme improves the network efficiency and further decreases the latency caused by large number of hop counts.



(a)



(b)

Figure 6. (a): Network latency for 16 cores, (b): Network latency for 64 cores

Overall, the above experimental results verify that the proposed algorithm obtains less power consumption and lower network latency while maintaining better performance compared to previous studies. On one hand, leveraging short paths between the subnets is able to decrease the number of hotspots, since most routers hold more ports compared to the root ones, packets would obtain more choices for paths. On the other hand, the power consumption is directly reduced with less routers.

VII. CONCLUSIONS

The architecture design plays an important role in the performance of NoC applications. In this work, we propose a SMoT architecture along with a path-wise scheme which takes into account both network topology and the congestion coefficient. Experimental results demonstrate that our proposed model obtains better performance while reducing the power consumption and network latency effectively and efficiently.

REFERENCES

[1] Manna, K., Chattopadhaya, S. Sengupta, I., An efficient routing technique for mesh-of-tree-based NoC and its performance comparison. International Journal of High Performance Systems Architecture, v4, n1, pp. 25-37, 2012.

[2] Lizhong Chen et al., NoRD: Node-Router Decoupling for Effective Power-gating of On-Chip Routers. Proc.

Of IEEE/ACM 45th International Symposium on Microarchitecture, MICRO 2012, pp. 270-281, 2012.

[3] Ghosal, Prasun, Das, Tuhin Subhra, FL2STAR. A novel topology for on-chip routing in NoC with fault tolerance and deadlock prevention. 2013 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT2013, 2013.

[4] Viswanathan et al., An optimised 3D topology for on-chip communications. International Journal of Parallel, Emergent and Distributed Systems, v29, n4, pp. 346-362, 2014.

[5] Michihiro Koibuchi et al., A Case for Random Shortcut Topologies for HPC Interconnects. 2012 39th International Symposium on Computer Architecture, ISCA 2012, pp. 177-188, 2012.

[6] Ebrahimi et al, LEAR-A low-weight and highly adaptive routing method for distributing congestions in on-Chip Networks. 20th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2012, pp. 520-524, 2012.

[7] Ezz-Eldin, Rabab, Magdy A. El-Moursy, and Amr M. Refaat. "Low leakage power NoC switch using AVC." Circuits and Systems (ISCAS), 2012 IEEE International Symposium on. IEEE, 2012.

Research on the Application of XML in Fault Diagnosis IETM

Zhiliang Zheng
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 15797708414@163.com
*corresponding author

Bailin Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 498194312@qq.com

***Abstract*—Fault diagnosis IETM is a digital technical manual which integrates editing, management, display and publication and faces the field of fault diagnosis. In its development process, there are some problems such as complex data management and difficult interaction. The successful application of XML technology becomes the key to solve these problems. In this paper, based on S1000D IETM development standard, the main characteristics of XML and its related technologies are first introduced, and then the application of XML and its related technologies in fault diagnosis IETM is analyzed and studied, including the creation of fault information data module, the display control of fault information and the comprehensive management of data. A conversion method between XML document data and database data is proposed, and the design process of fault diagnosis IETM based on XML is discussed. Finally, a fault diagnosis IETM display system based on XML is preliminarily developed by combining the related technologies mentioned above. Experiments show that XML technology can effectively manage the information in fault diagnosis IETM and display it to users in an interactive way, which has important theoretical value and certain practical significance for developing fault diagnosis IETM and exploring the deeper application of XML in IETM.**

Keywords-XML; S1000D; Data Module; Fault Diagnosis; IETM

I. INTRODUCTION

Interactive electronic technical manual (IETM) is a digital technical manual made according to the standard digital format and in the form of text, graphics, tables, audio and video [1], which provides fault diagnosis, maintenance and other functions for weapon equipment or civil equipment through human-computer interaction [2]. Fault diagnosis IETM is an integrated IETM system focusing on the field of equipment fault diagnosis. The most basic work of IETM is to process technical data. How to manage a large amount of data information? How to use the stored information for fault diagnosis? How to use information to complete the interaction with users? The effective solution of these problems is the key to develop fault diagnosis IETM. In the common technical framework of IETM shown in Figure 1, XML and its related technologies account for a large proportion. It can be seen that XML technology plays a very important role in IETM development. This paper takes the above three problems as the background and according to the international standard S1000D standard as the IETM standard to describe the application of XML in fault diagnosis IETM, including using XML to create fault information data module, displaying control equipment fault information and using XML to build IETM database. It systematically expounds the relationship between XML Technology and fault diagnosis IETM, so as to provide solutions and ideas for the above problems.

DOI: 10.2478/ijanmc-2022-0023

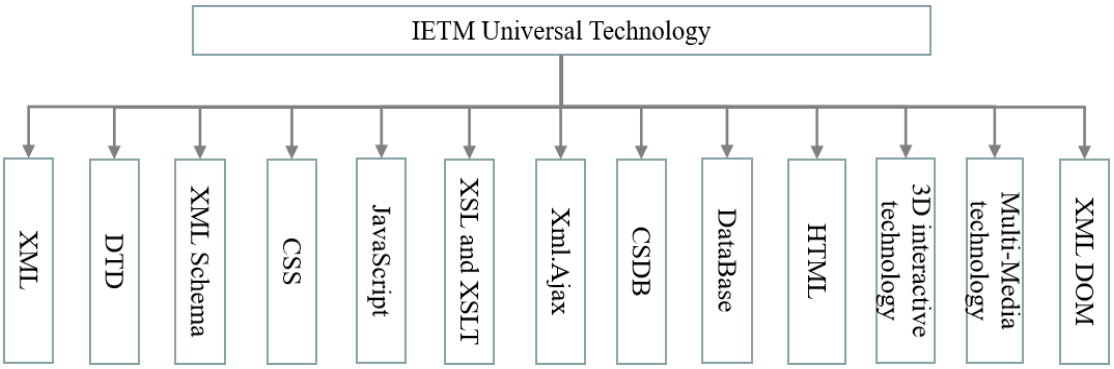


Figure 1. Technical framework commonly used by IETM

II. Summary Of XML And Its Related Technologies

A. Summary of XML Technology

Extensible markup language (XML) [3-5] is a meta markup language. It is an effective tool designed to transmit and store data depending on some rules and applications. It is flexible, simple and easy to expand. Users can use any tag to define XML instance document [6]. An XML instance document contains elements and attributes. The document body has three parts: preface, body and end, which are respectively composed of multiple element nodes. Each element node has its own attributes and can expand child nodes. The writing of XML document is essentially to organize each element node reasonably. XML integrates data reuse, data display separation, scalability, syntax freedom and structured integrated data. It has unique advantages in data storage and management, powerful super connection, cross platform, cross network, cross program language data description, efficient organization of a large amount of data and so on. Of course, XML also has some disadvantages. For example, it allows users to develop tags with different rules according to needs. Although it makes the search easier and flexible, it also increases the complexity of documents, and the same data will occupy more space than stored in other style databases [7]. Therefore, when applying XML to IETM, we should take comprehensive consideration and reasonably define tags.

B. Overview of XML constraint rules

XML documents need to follow several basic rules when editing, including having and only one root element node; If each element node has a start tag, it must have an end tag; Each element node should be reasonably contained in the parent element and cannot be cross referenced; In order to make the creation of an element meaningful, the attribute of an element cannot be empty. In addition to the basic rules that need to be followed, when creating XML documents, appropriate schemas should be selected to describe and restrict their document structure. For this purpose, XML provides two schemas: document type definition (DTD) and XML schema. DTD [8] is a set of syntax rules about markers. It defines the logical structure of the document and specifies the elements, entities, attributes of elements and the relationship between elements and entities used in the document. XML schema [9] is the successor of DTD. It improves the shortcomings of DTD, such as non-extensibility, non-compliance with XML syntax, non-support for namespace applications, support for a few simple data types, etc. XML schema is the standard schema recommended by W3C at present, and is gradually becoming the mainstream schema of XML description and binding document structure. The analysis and comparison of the two are shown in Table I. Through the analysis and comparison of the two, XML schema itself provides a more standardized and complete mechanism to restrict an XML file than DTD, which has obvious advantages. Therefore, in the description process, this paper selects XML scheme schema to describe and restrict the XML document structure.

TABLE I. TABLE TYPE STYLES

Technical model	Main advantages and disadvantages
DTD	DTD comes from and is a subset of SGML. It is the standard before schema and an important part of XML1.0. Its definition is a specific application domain. It is mainly used for validation and good form verification. The main advantage of DTD is that it can quickly determine whether a document is valid by using a parser to match regular expressions and data patterns in the document. Its main disadvantages are that it uses non-XML syntax, requires special processing tools, and does not support various data types and spatial mechanisms. DTD cannot introduce other DTDs, and its scalability is poor.
XML Schema	XML Schema is the standard recommended by W3C at present. It introduces data type and namespace, supports internal reference of schema, has good expansibility, can be converted between different schemas through mapping, and has good data interchangeability. For this trend, this schema will replace DTD and be used to define rules for XML documents in all types of network applications. Its main advantage is that the schema itself is an XML document with good data scalability, consistency and accuracy. From the user's point of view, it has better simplicity, legibility and ease of use, better support for data exchange and stronger data binding. The main disadvantage of is that the schema cannot define entities like DTD, and it is more complicated than DTD.

C. XML DOM overview

XML DOM, fully known as XML document object model, is a standard model for XML documents. It defines the standards and methods of accessing and operating XML documents. XML DOM regards the whole XML document as a node tree, in which each component is defined as a node that can be accessed and processed. In the XML DOM node tree shown in Figure 2, each node is independent and related to each other.

For example, when the user accesses the XML element and the attributes of any node in the XML document, the user can access the XML element or delete the attributes of any node through JavaScript. XML DOM is a powerful tool for users to operate XML documents. It can not only arbitrarily control the content of the whole XML document, which brings great flexibility to the development of fault diagnosis IETM, but also plays an important role in the integration and transformation between relational database and XML [10].

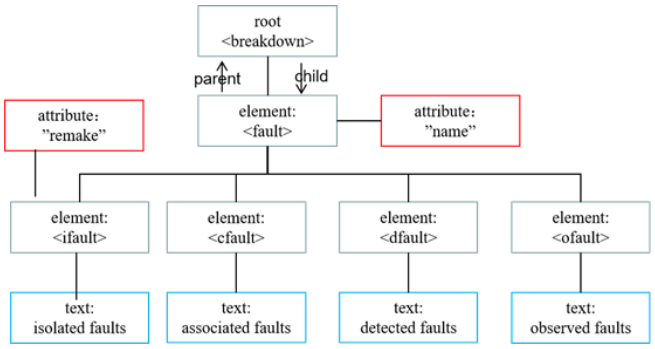


Figure 2. XML DOM tree example

III. XML IS APPLIED TO THE CREATION OF DATA MODULE

A. Data Module

IETM includes various technical information such as equipment function, performance, operation principle, program, operation, maintenance, detection, inspection, fault report and exception handling, disassembly procedure, assembly... Which need to be stored [10], and then display pictures, words and other information on the electronic screen for users to consult and browse [11]. S1000D standard stipulates that technical information is organized by data module (DM), which is the smallest information unit in IETM technical data. S1000D standard defines 13 types of DM, as shown in figure 3, among which 8 types of DM are commonly used.

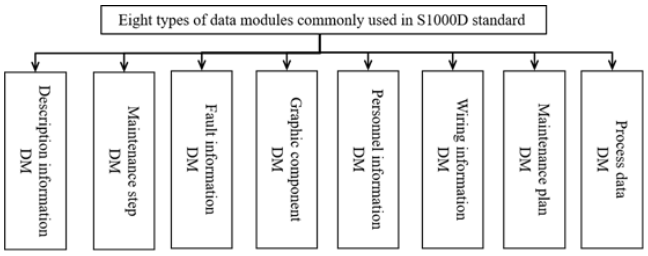


Figure 3. Eight types of DM commonly used in S1000D standard

What may be needed in fault diagnosis IETM include description information data module, maintenance step information module, fault information data module, maintenance plan data module, etc. The data module consists of two parts: Identification and status segment(IDSTATUS) and

Content segment(CONTENT). The Identification and status part is used to identify and control the data module, which is mainly used to identify and locate the data module. This part is invisible to the user. The data module code (DMC) and the version number of the data module constitute the unique identification of the data module, which is used to manage the data module in CSDB. The content part describes most of the information that users can view in the data module. This paper mainly takes the creation of fault information data module as an example to describe the role of XML in the creation of data module.

B. Fault information data module

The fault information data module is an information collection that gives specific fault finding and troubleshooting methods according to different fault causes during the design and use of the equipment. It is used to describe the phenomenon, cause, relevant information of fault parts, brief solutions or detailed fault isolation procedures. It is the main basis for the fault diagnosis IETM system to realize the fault diagnosis function. The fault information data module is divided into three parts: reference information, fault report information and fault isolation procedure. Among them, the fault report describes the fault content, fault isolation describes the troubleshooting of the fault, and the reference information is used to quote other technical data. The constituent elements of each part of the fault information data module are shown in Figure 4.

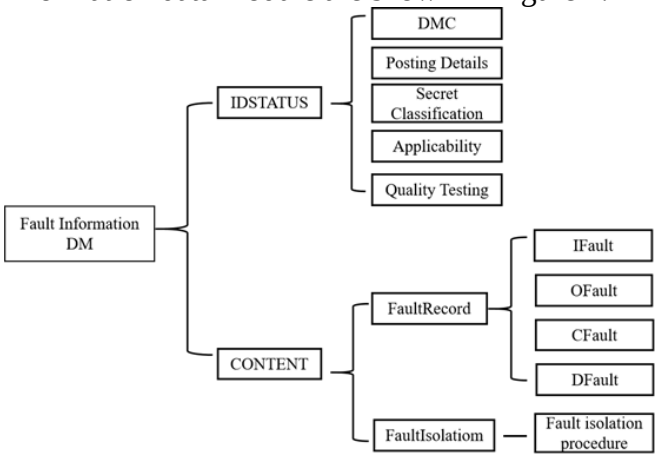


Figure 4. Structure diagram of fault information data module

C. Creating fault information data module with XML

In order to meet the principle of "Edit once, reuse" and avoid the uncertainty of user-defined labels, when creating the fault information data module, XML stipulates that there are specific element labels to describe the information of each part of the module, in which the fault report is divided into the following four categories: isolated faults described by element <ifault>; Detected faults described by element <dfault>; Observed faults described by element <ofault>; Associated faults described with element <cfault>. Fault isolation information describes the fault information and gives the basic steps of fault isolation, which is described by element <afi>. Fault isolation information includes various information elements describing fault isolation diagnosis. This part is the focus of studying equipment fault information model and the inevitable factor to improve the success rate of fault reasoning task [12]. The final generated data module is a complete XML file. Figure 5 shows the breakdown.xml of fault information data module document fragment.

```
<?xml version="1.0" encoding=" UTF-8"?>
<breakdown><name>GJB6600BIKE</name>
    <ifault><issno issno="001" type="string"><information>Isolated fault</information></issno></ifault>
    <dfault><issno issno="002" type="string"><information>Detected faults</information></issno></dfault>
    <ofault><issno issno="003" type="string"><information>Observed faults</information></issno></ofault>
    <cfault><issno issno="004" type="new"><information>Associated fault</information></issno></cfault>
    <afi>Fault isolation information</afi>
</breakdown>
```

Figure 5. Code snippets for XML documents

When creating the fault information data module, this paper selects the XML scheme mode for breakdown Generate file breakdown.xsd, Figure 6 is the description form of fault information data module in schema mode.

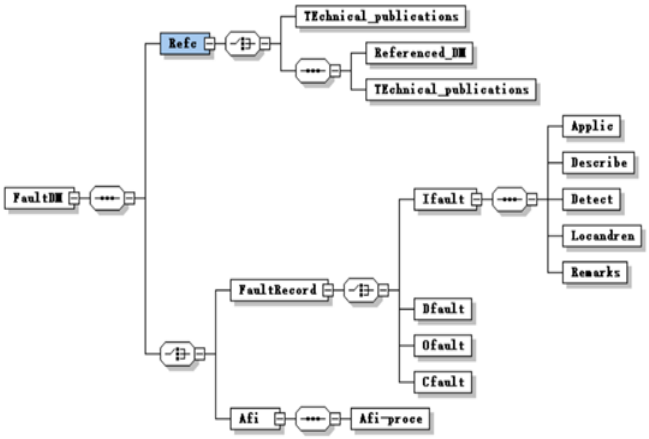


Figure 6. Schema mode of fault information DM

When using, you need to add a pattern reference to the header of the generated fault information data module:

<xmlns:xsi="http://www.w3.org/2001/XMLSchemainstance"xsi:schemaLocation="http://www.w3schools.com breakdown.xsd">

Because XML schema itself is an XML document, and the basic syntax rules of defining document structure through XML are consistent, the created data module can be shared and used in multiple IETM systems.

IV. XML DISPLAY CONTROL FAULT INFORMATION

The display control of fault information in XML mainly includes two parts. One part is the storage and query of fault information, that is, the information is classified and written into the XML document by using the tag, so that the fault information has certain structural relationships such as inclusion, connection and dependence through the element tag. At this time, the XML document will be modeled as an annotated diagram, which is actually a tree sequence, and each tree represents a document or element [13]. When making IETM system, data module code, information control code, information type and metadata information are written and stored in XML. For example, the creation of fault information data module is the process of information storage; Query is to find out the fault information that meets the needs of users in the data module, that is, the XML file set, according to the relevant fault query conditions submitted by users. Due to the increasing popularity of XML for data representation, people query XML mainly through keyword query [14], which needs to access the structural tags in XML to obtain the content. The other part is to provide technical support for the interactive fault diagnosis process by describing the fault reasoning process in the data module, and realize the information interaction by controlling the label of XML. In the whole process of XML display and control of fault information, it also involves the transformation with database.

A. XML display style

Because XML documents can customize labels, have no information to display data, and separate data from display styles, corresponding technologies are needed to display information to users. XML can be displayed using Cascading style sheet (CSS), Extensible style sheet language transformations (XSLT) and scripting language JavaScript. For the breakdown.xml in Chapter 2 uses three methods as follows.

1) Using CSS style sheets: Cascading style sheet (CSS) is a computer language used to represent file styles such as HTML or XML. It can not only modify the web page statically, but also format the elements of the web page dynamically with various scripting languages. There are two ways to use CSS to control the display format of XML. One is to directly use CSS style by embedding <html: style> in the XML document. You must first define the namespace of HTML like "http://www.w3.org/TR/REC-html40" before you can use the tags in HTML The syntax is shown in Figure 7.

```
<element xmlns: html = "http://www.w3.org/TR/REC-html40">
  <html: style>
    ifault {background-color: #ffffff; width: 100%;}
    dfault {background-color: #ffffff; width: 100%;}
    ofault {background-color: #ffffff; width: 100%;}
    cfault {background-color: #ffffff; width: 100%;}
  </html: style>
</element>
```

Figure 7. XML document embedded CSS example

The second is to break down the tags in the XML file CSS file. Some code examples of CSS file are as follows:

ifault {background-color: #ffffff;width: 100%;}

Then link the CSS file to the XML file, so that the tags in the XML document can display with breakdown.css file.

<?xml-stylesheet type="text/css" href="breakdown.css"?>

Compared with the first method, this method is more flexible and convenient. It can not only apply CSS files to multiple XML documents, but also write the label styles of multiple XML documents in one CSS file to realize one-time generation and reuse.

2) Using XSLT style sheets: XSLT is a language used to transform one XML document into another XML document or other types of documents. The principle is to convert the XML file into a document that can be recognized by the browser, such as HTML, before the browser displays it. When using it, you need to generate breakdown.xsl file, The code fragment is shown in Figure 8.

```
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:template match="/">
    <html>
    <body style="font-family:Arial;font-size:12pt; background-color:#EEEEEE;">
        <h2>Breakdown</h2>
        <xsl:for-each select="ifault">
          <xsl:value-of select="title"/>
          <xsl:value-of select="artist"/>
        </xsl:for-each>
    </body>
    </html>
  </xsl:template>
</xsl:stylesheet>
```

Figure 8. Code snippets for XSLT documents

And through the breakdown.xml add The following code:

<? xml-stylesheet type="text/xsl" href="breakdown.xsl"?>

reference XSLT style sheets to XML documents to work. Through breakdown.xsl file, you can choose to add elements and attributes to the output file, remove elements and attributes, rearrange and classify elements, test and hide or show which elements.

3) Use JavaScript to dynamically display: This method mainly uses the XMLHttpRequest object to parse the XML document into the XML DOM object, and then operate the DOM object through JavaScript to obtain the information in the XML document and display it. The specific method is to define a new XMLHttpRequest object through the xmlhttp = new xmlhttprequest() operation. When the system adopts the server centered B / S structure, the XMLHttpRequest object can be used to exchange data with the server in the background, and then send a request to the server using the open() and send() methods of the XMLHttpRequest object. The method is xmlhttp. open("GET","breakdown.xml",false);xmlhttp.sen d(); "breakdown.xml" in the open () method is the address of the XML document on the server, and finally xmldoc = xmlhttp.responseXML; Use the responsexml attribute to return the XML document object. At this time, you can use JavaScript to manipulate the methods and attributes of the DOM node tree to check and parse the object. For example, x = xmldoc getElementsByTagName("ifault"); Locate the tag name in the XML document through the getElementsByTagName () method to obtain the content in the tag.

In the above three methods, general XSLT is the preferred XML style sheet language, and its display control flow is shown in Figure 9. During display, the required data module and its corresponding XSLT style are sent to the reader for display, and the XSLT style written in a certain mode has certain universality and can be reused.

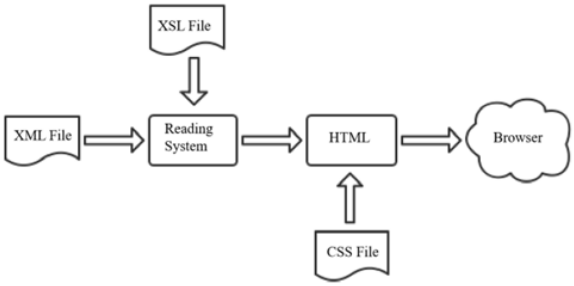


Figure 9. XML display control process

B. XML oriented interactive fault reasoning mechanism

The interactive fault reasoning mechanism can adjust the user's fault symptom input as the premise part of the diagnosis rules, and then match the knowledge base. If the matching is successful, the action part of the rules will be executed until the diagnosis conclusion is reached or the rule

base is exhausted. Or write the internal logical relationship of the question into the XML document, analyze, reason and judge the user's answer to the previous question to put forward the next question, and get the final diagnosis conclusion in this cycle. In the above two mechanisms, the diagnostic rules, knowledge base information and problems involved in reasoning can be described in the form of XML and displayed in the style in Chapter 3.1 when necessary.

C. Relationship between XML and database

There is a core concept in S1000D standard - common source database (CSDB). As a container for storing information and management information in IETM system, CSDB stores all the information required by the technical manual, including data modules. The data modules mentioned above are composed of XML documents, so that XML documents can be accessed through the database. At the same time, pictures, videos, audio and other data are also stored in the database, and the information in the database can be referenced in the XML document, Therefore, it is very important to select an appropriate mapping model from XML to database [15-16] to realize the mutual transformation between XML documents and database data. Using table mode, there are two main forms of conversion between XML document and database.

1) The transformation from XML document to database: parses the label in XML document into the form of table and inserts it into the database. Starting from the root node in XML document, take the label information as the attribute in the table until the last label in XML document is traversed. The attribute field definition of each tag in the database is shown in the table Ⅱ. To take the <breakdown> tag in the breakdown.xml document as an example. The XML document number starts from 1 and the tag number starts from 1. Then the fields in the <breakdown> tag corresponding table have XmlID = 1, LableID = 1, LabelName = "breakdown". At the same time, two tag types are defined in XML, one is the element tag element and the other is the attribute tag attribute. Here, breakdown is the element tag, so LabelType = "element". The tag does not contain text information, and LabelVal = null. Because < breakdown > is the root node, there are no parent tags, but only child tags, so LabelParent = 0 and LabelChild = 1. By analogy, in this way, the label data in the XML document is mapped to the data in the database for storage, and the information in the XML document can be obtained by accessing the table from the database.

TABLE II. LABEL ATTRIBUTE FIELD DEFINITION

Attribute field	Data type	Field description
XmlID	Integer	XML document ID of the tag
LabelID	Integer	ID of the tag
LabelName	Varchar(30)	The name of the label
LabelType	Varchar(30)	Type of label
LabelVal	Varchar(255)	Text information in label
LabelParent	Integer	The ID of the label's parent label
LabelChild	Integer	The ID of the child tag of this tag

2) The mapping from database information to XML document: That can create an XML document for each table in the database, take the table name as the root node of the XML document, generate the sub elements under the root node with the information in the table in behavioral units, and a row of information is a sub element node, and then associate the fields in each row of information as the sub elements or attributes of the sub nodes, and the attributes of the fields as the attributes of the sub element labels. When XML schema is used to constrain the structure of XML document, it shall be consistent with the field type defined in the database. For example, if the data type of field x defined in the database table is integer, only numbers shall be used for the text of label <x> in XML document. In this way, a label node is generated in the XML document until the last field of the last row of information in the table.

V. EVELOPMENT PROCESS AND APPLICATION IMPLEMENTATION OF IETM FOR FAULT DIAGNOSIS BASED ON XML

XML can store data according to needs, transmit content in an interactive way, and provide the interface between data and external elements to operate data. It realizes the efficient storage and

sharing of information between different systems and reduces the development cost. Therefore, XML is the best choice for developing fault diagnosis IETM[17]. Taking the IETM in XML format generated from traditional paper documents as an example, from data storage to display to users, and then to fault diagnosis in an interactive way, the design flow of XML based fault diagnosis IETM is shown in Figure 10.

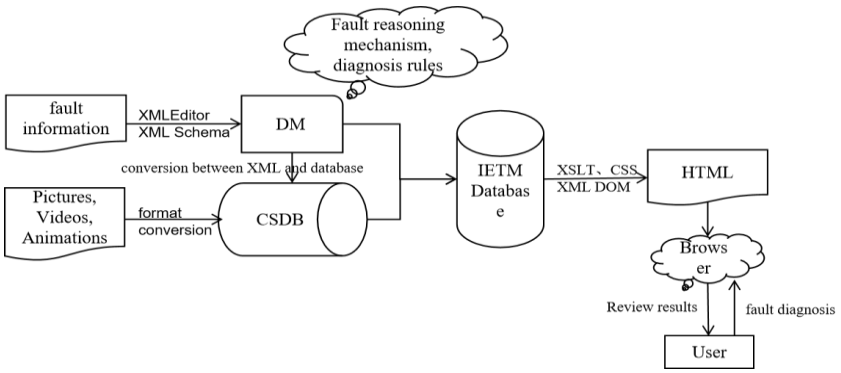


Figure 10. Design flow of fault diagnosis IETM based on XML

Firstly, the fault information is analyzed from data storage to display, and it is allocated to the corresponding data module through XML editor according to the modular design idea. Then, the structure of each data module is described and constrained based on XML schema; Then, the XML document is stored in the database according to certain conversion rules; Finally, the display style sheet of XML document is written, and the data is retrieved and output to the browser according to the needs of users by converting it into HTML document. Figure 11 is an interactive fault diagnosis interface for users. In this interface, users can not only consult the displayed fault information for diagnosis, but also input fault information, complete the diagnosis by using the fault reasoning mechanism and diagnosis rules stored in the data module, and obtain the diagnosis results and relevant information through the above displayed process. The specific implementation methods of each part have been described in the previous chapters of this paper, and will not be described here.

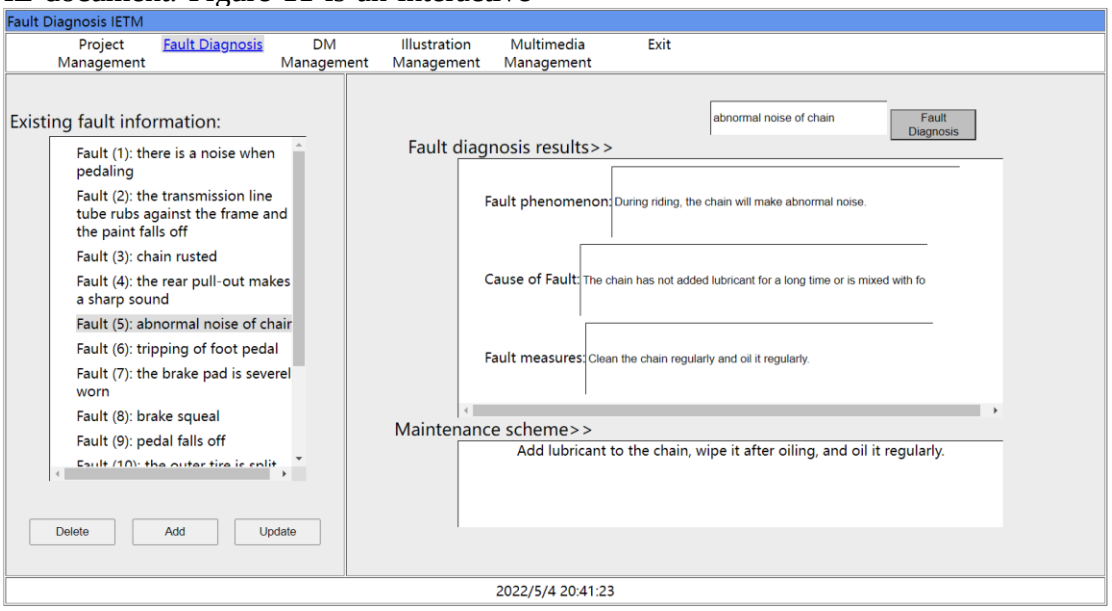


Figure 11. Fault diagnosis interface

VI. CONCLUSION

XML technology provides an effective means for efficient storage, integration and sharing of data in fault diagnosis IETM. From the creation of fault information data module, to the management and display of fault information, to the association with database, XML covers almost every stage of IETM development. This paper systematically analyzes the application of XML technology in fault diagnosis IETM, and expounds the technologies involved in each application process in detail. By integrating these technologies, the preliminary display system of IETM for fault diagnosis is realized. The research shows the key role of XML technology in IETM development, which provides ideas for developing IETM based on XML in the new situation, and is of great significance to system development in the field of equipment maintenance and support and equipment fault diagnosis. At present, the research on XML technology and the implementation of the system are still in the initial stage, and the follow-up work should consider the combination with other advanced technologies to further explore the deeper application of XML technology in fault diagnosis IETM.

REFERENCES

[1] Sun H B, Xu Z C, Zhou J. Research and Design of the Remote Fault Diagnosis System for Complicated Equipment Based on Intelligent IETM [J]. Advanced Materials Research, 2012, 490-495:1564-1568.

[2] Yaguo, Lei, et al. "Opportunities and Challenges of Machinery Intelligent Fault Diagnosis in Big Data Era." Journal of Mechanical Engineering (2018).

[3] Extensible Markup Language (XML). https://www.w3.org/TR/xml/, https://www.w3.org/TR/xml11/

[4] Banzal S. XML Basics [M]. Mercury Learning & Information:2020-08-31.

[5] Deng Jia, Liu Hongxing. Research and Development of a XML Modeling Tool[C]//.Proceedings of 2018 17th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES 2018）, 2018:337-340.

[6] Zhang Fu, Li Qiang. Constructing ontologies by mining deep semantics from XML Schemas and XML instance documents [J]. International Journal of Intelligent Systems, 2021, 37(1).

[7] Liang W J, Jia J N, Cai L Y, et al. A Survey of Common Technology and Data Model Standard of Equipment IETM [J]. Advanced Materials Research, 2014, 998-999:553-557.

[8] Tae Gwon Kim. A Study on Processing XML Documents [J]. Journal of KIISE, 2016, 43(4).

[9] Dongyang Liang, Shasha Li, Jie Yu, Bin Ji. Design and Implementation of an XML Schema Based XML Data Editor[C]//.Proceedings of 2019 the 9th International Workshop on Computer Science and Engineering (WCSE 2019). [Publisher unknown], 2019:663-669.DOI:10.26914/c.cnkihy.2019.038272.

[10] Husam Ahmed Al Hamad. RXML: Path-Based and XML DOM Approaches for Integrating Between Relational and XML Databases [J]. International Journal of Database Management Systems, 2017, 9(5).

[11] Jiaju Wu, Zhenji Liu, Xinglin Zhu, Rong Yu. The Study on Lucene Based IETM Information Retrieval[C]//Proceedings of 2016 International Conference on Communications, Information Management and Network Security (CIMNS2016), 2016:230-233.

[12] Lirong Meng, Yongqi Ma, Xun Cheng. Research on Interactive Fault Reasoning of Complex Agricultural Equipment by IETM [J]. IOP Conference Series: Earth and Environmental Science, 2020, 474(3).

[13] Jan Hidders, Jan Paredaens. A Formal and Unified Description of XML Manipulation Languages [J]. Fundamenta Informaticae, 2016, 145(4).

[14] Bai Luyi, Cui Zengmei, Duan Xinyi, Fu Hao. Keyword coupling query of spatiotemporal data based on XML [J]. Journal of Intelligent & Fuzzy Systems, 2022, 42(3).

[15] Model-based XML to Relational Database Mapping Choices [J]. International Journal of Recent Technology and Engineering, 2019, 8(3S).

[16] Sabirovna K M. Application of xml technology for data design [J]. ACADEMICIA An International Multidisciplinary Research Journal, 2020, 10(6):1681.

[17] Sheng W P, Wang R Q, Lin Y, et al. Application and Research of XML Technology in IETM [J]. Advanced Materials Research, 2012, 424-425:1062-1064.

Research and Implementation of Image Rain Removal Based on Deep Learning

Dong Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail:queerwd@163.com

Zhongsheng Wang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail:wzhsh1681@163.com

***Abstract*—The traditional rain removal algorithm needs to optimize a large number of parameters, and it is only effective for rain of a specific shape, and the model generalization ability is poor. In recent years, the performance of rain removal methods based on deep learning is better than many traditional methods, but there are problems such as incomplete or excessive rain removal, and incomplete texture reconstruction of background details. This paper proposes a rain removal network based on generative confrontation, which connects the high and low frequency parts and integrates them into the model. At the same time, the attention mechanism cyclic neural network is organically combined, which can better preserve the background texture while removing rain. Theoretical can produce better rain streak removal with better color distortion.**

Keywords-Deep Learning; Generative Confrontation; High and Low Frequency; Attention Mechanism; Background Texture

I. INTRODUCTION

The network design motivation and method are briefly described as follows. The algorithm mainly includes three parts: frequency decomposition module, generator network [1], and discriminator network. These three parts will be introduced separately below.

A. Frequency decomposition module

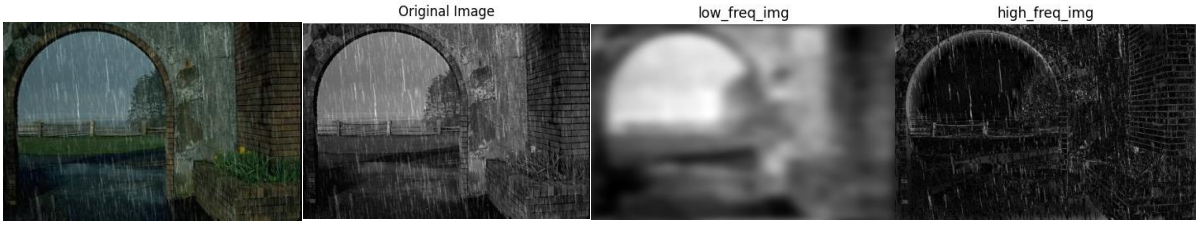
When training the mapping from rain images to no rain images directly on the entire image domain, the mapping range covers all possible pixel values, and the rain trace information and background information in the image will be highly aliased, making it difficult to combine the rain trace information with the background information. The information is accurately separated, causing the background information of the reconstructed image to be blurred, and there are still some rain marks left, and with the deepening of the network depth, the phenomenon of gradient disappearance may occur. Considering that almost all the information of rain marks exists in the high-frequency part of the image, the image with rain and the image without rain are decomposed into two parts: high-frequency layer and (detail layer) and low-frequency layer (base layer) respectively through guided filters, as shown in the formula: (1) and (2). Compared with the traditional bilateral filter, the use of the guided filter to decompose the image has the advantages of higher computational efficiency and stronger edge protection characteristics, and the guided filter is also more accurate in the processing of image details in terms of filtering effect.

DOI: 10.2478/ijanmc-2022-0024

$$X = X_{base} + X_{detail} \quad (1)$$

$$Y = Y_{base} + Y_{detail} \quad (2)$$

Figure 1 shows an example of image frequency decomposition. It can be found by observation that the edge contour information and rain trace information of the image are preserved in the detail layer of the image with rain [2], while the base layer of the image with rain is similar to the base layer of the image without rain. , only there is a difference in the image detail layer.



(a) Image with rain (b) Image with rain (c) Low frequency map of rain images (d) High frequency map with rain

Figure 1. Example of Frequency Decomposition

It can be seen from the comparison results that the rain image is decomposed into high-frequency regions and low-frequency regions according to the frequency, and most of the rain marks exist in the high-frequency region. The low frequency map is integrated into the generator as an additional constraint for better results. The frequency decomposition module is shown in Figure 2.

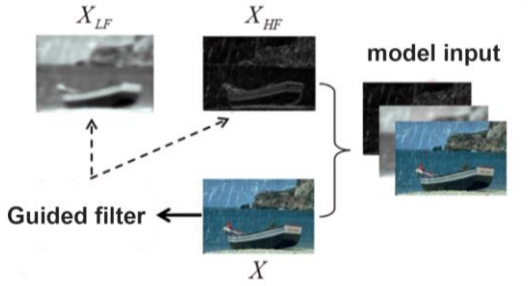


Figure 2. Frequency Decomposition Module

B. Generator network

The generator network proposed in this paper is shown in Figure 4 and is mainly divided into four parts: (1) a convolutional layer receives the input image; (2) the combination of several residual blocks ResBlocks is used to extract deep feature information; (3)The recurrent unit adopts the long short-term memory unit LSTM, and this stage takes the output of the input and the state of the recurrent layer of the previous stage as the input; (4) A convolutional layer outputs the rain-removed result image. Represented by the following formula:

$$x^{t-0.5} = f_{in}(x^{t-1}, y) \quad (3)$$

$$s^{t} = f_{recurrent}(s^{t-1}, x^{t-0.5}) \quad (4)$$

$$x^{t} = f_{out}(f_{res}(s^{t})) \quad (5)$$

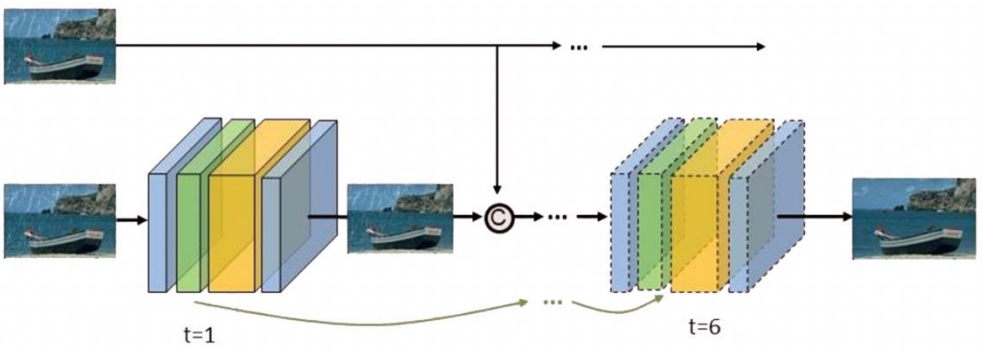


Figure 3. The network structure of the generator

- All convolution kernel size=3×3 and padding=1×1, and the activation function uses ReLU. In the whole process of the network, there is no up-sampling operation, and other settings of the convolution kernel are designed to keep the resolution of the feature map unchanged, and the restored clear image will not lose the details of the image content. The first convolution kernel consists of a Conv+ReLU; the residual block consists of 5 ResBlocks, each ResBlock consists of two Conv+ReLU and an attention mechanism module; the last convolution layer consists of a Conv, no activation function.
- MSE loss: Calculates the output of the last stage (for example, the last optimization stage is T) and the MSE of the ground truth.

$$L_{MSE} = \| x^T - x^{gt} \|^2 \tag{6}$$

- Negative SSIM loss: Calculates the output of the last stage (such as the last optimization stage is T) and the ground truth SSIM.

$$L_{\text{SSIM}} = -SSIM(x^T, x^{gt}) \tag{7}$$

C. Discriminator network

In the generative adversarial network [3], the generator is responsible for generating the corresponding data, the purpose is to "fool" the discriminator, the role of the discriminator is to judge whether the input data is real or generated, and the purpose is to find out the "fake data" generated by the generator. Through continuous training, the ability of the discriminator of the generator is getting stronger and stronger, and the final generator can better realize the corresponding network function. Generative adversarial networks have great potential because they can learn and simulate the distribution of any data.

As mentioned above, although the network function is implemented by the generator, and only the generator is used in the testing process, the quality of the discriminator is crucial to the performance of the generator. If the discriminator is not set properly, after a small amount of training, the "fake data" generated by the generator is enough to make the fake data look real, and the discriminator cannot judge the real and fake data. At this time, although the generator's data is still "bad", due to the referee's standard is low, it is easy to produce "proud" mentality, and the performance cannot be further improved.

In standard generative adversarial networks, the discriminator network distinguishes whether the

input image is a real image or a fake image generated by the generator by performing binary classification, and the generator is trained to generate a fake image to convince the discriminator that it is real. Whereas in relative generative adversarial networks, the discriminator is also trained to reduce the probability that a real image is real, and the discriminator estimates the probability that a real sample is more real than a fake one.

II. EXPERIMENTAL SETUP

A. Existing Synthetic Rain Image Datasets

Since it is very difficult to obtain a large number of rain images and corresponding rain-free background images from the real world, this paper uses five commonly used synthetic benchmark rain image datasets [6], Rain100L, Rain100H, Rain800, Rain12600 and Rain12 to train and evaluate the proposed algorithm. Rain Network[4].

B. Real Rain Image Datasets

Internet-Data: Contains 149 real rainy images in total, and Figure 7 shows three typical real rainy scene images.

C. Production of scene-based depth rain image dataset

Most of the existing synthetic rain image datasets and rain removal methods are implemented based on the rain image model of Equation (8):

$$I(x) = B(x) + R(x) \tag{8}$$

Where $B(x)$ is the clean background image without rain, $R(x)$ is the additional rain streak image, and $I(x)$ is the composite rain streak image. In heavy rain conditions, where the accumulation of rain streaks can cause attenuation and scattering, the visibility is spatially different in the image, resulting in a fog or "veil" effect in the captured image, and when the depth of the scene varies greatly in the rain image, the visibility of objects will change with the scene depth of the image, distant objects are more visually occluded by fog, and when the rain is heavy, the occlusion effect is more obvious.

This paper uses the depth information of the image to synthesize a rain data set that is more in line with the real rain scene based on the rain mark image designed in this paper and the construction site data set as the background [7]. The rain mark images provided in this paper take into account the information such as the density, direction, and scene depth of the rain marks [8]. A total of 400 images of construction sites are selected for rain image synthesis. Each image uses the depth information and rain marks in different directions to generate 12 buildings. A total of 4800 images are synthesized for the rainy day images of the construction site, of which there are 4500 images in the training set and 300 images in the test set. This method makes the rainmarks of the synthetic images more diverse and provides a more realistic rain image dataset for training the rainmark removal network.

D. Evaluation indicators

It is not objective and accurate to evaluate the quality of image restoration only through the visual perception of human eyes, because different human eyes have slightly different image resolution capabilities, which are greatly affected by subjective perception, and are also easily affected by external objective factors. Therefore, it is very important to use appropriate image evaluation indicators to accurately and objectively evaluate the effect of image rain streak removal. In

the field of image rain removal, researchers use data quantification of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) to measure the quality of image rain streak removal. PSNR and SSIM will be introduced separately below.

PSNR is mainly aimed at the absolute error between the corresponding pixels, and it does not fully consider the visual characteristics of the human eye. Therefore, when evaluating the quality of image reconstruction, the evaluation results are usually inconsistent with human subjective visual perception. . Compared with PSNR, SSIM is more in line with the human eye's judgment of image quality in the measurement of image quality.

In the image rain removal task, SSIM is an evaluation index used to measure the similarity between the restored rain-free image and the real rain-free image. The value range of SSIM is [0, 1]. The more similar the two images are, the closer the value of SSIM is to 1.

E. Experiment and result analysis

Comparative Methods: The proposed rain removal method is compared with the traditional optimization method GMM, as well as the deep learning based methods JORDER and RESCAN. The GMM uses a pre-trained Gaussian mixture model as prior knowledge to decompose the image into background scenes and rain marks. JORDER mainly proposes a multi-task joint detection rain removal network. RESCAN utilizes atrous convolution and residual learning for step-by-step rain removal.

In order to verify the effectiveness of the image rain removal method in this paper for image rain removal, this section verifies the effectiveness of the proposed method on a synthetic dataset through experiments, and compares it with GMM [9], JORDER, and RESCAN [10] for three representative images to remove rain. Algorithms were compared, and the experimental results were quantitatively analyzed using structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) evaluation indicators. It can be seen from the table 1 that the method in this paper has great advantages over other methods in both PSNR and SSIM indicators.

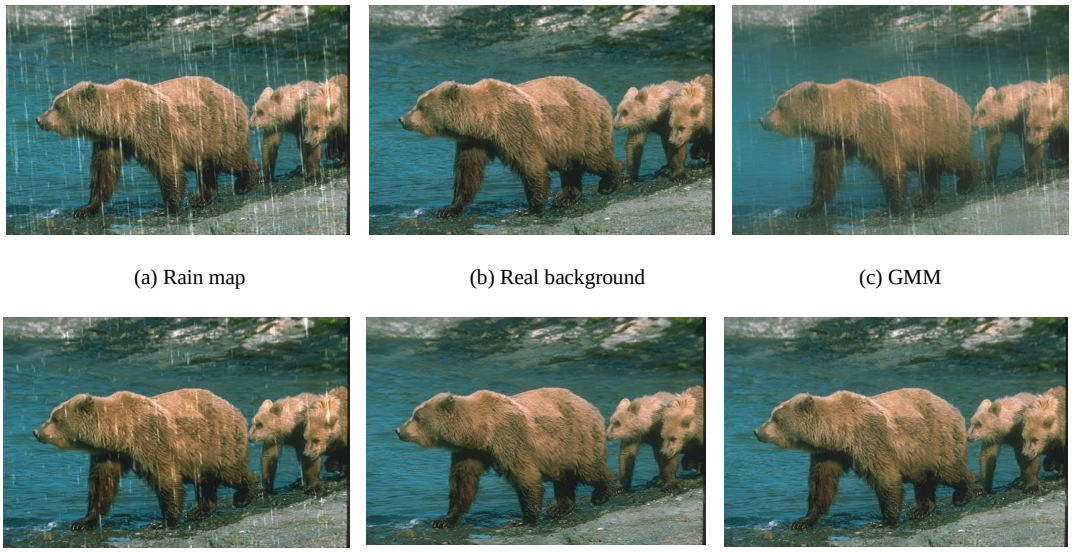
Figure 4 shows the different rain removal results respectively. Among them, sub-image (a) is the image with rain, sub-image (b) is the real image without rain, sub-image (c) is the GMM rain removal result, sub-image (d) is the JORDER rain-removing result, and sub-image (e) The result of draining for RESCAN, the sub-figure (f) is the draining result of this paper. They are described in detail below.

TABLE I. COMPARISON OF EXPERIMENTAL RESULTS ON SYNTHETIC DATASETS

Datasets	GMM	JORDER	RESCAN	Ours
PSNR	25.9725	26.4012	28.7863	29.1035
SSIM	0.9181	0.9246	0.9355	0.9661

In Figure 4, the amount of rain in the original rain image (a) is light rain, and the detail information covered by the rain marks is less, so more details can be added when removing the rain mark information. Although GMM can remove most of the rain marks, the pixel details of the image after rain removal are blurred; JORDER leaves a small amount of rain marks and makes the

background image darker; RESCAN and the method in this paper are very thorough in removing the rain mark information and restoring the background. The effect of removing rain marks is better.



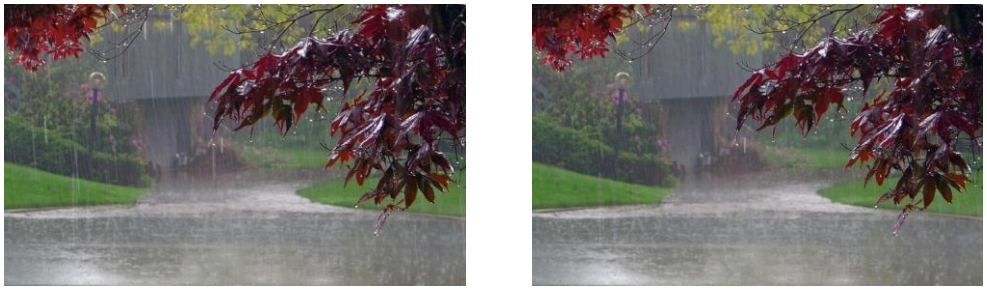
(a) Rain map (b) Real background (c) GMM

(d) JORDER (e) RESCAN (f) Method of this paper

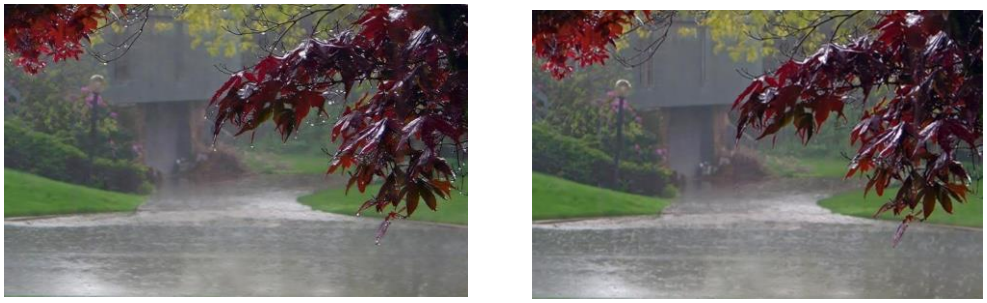
Figure 4. Visual comparison of the rain effect on the synthetic dataset Rain100H

F. Comparison on real rainy image datasets

The network model of the algorithm in this paper is trained on the synthetic rainy day data set[5]. Therefore, in the comparison of the real rainy day result images of the following methods, we will judge from various aspects according to human subjective vision, and continue to use two of the deep learning algorithms (JORDER and RESCAN) used in the comparison of synthetic datasets. Compared with the synthetic data set, the real rainy day situation is more complex, which makes the image quality of the picture more diverse than the synthetic rainy day data set, and can better detect the generalization ability of the rain removal method.



(a) (b)



(c) (d)

Figure 5. (a) Real rain map (b) JORDER (c) RESCAN (d) This paper method

The first image in Fig. 5 has less rainfall and residual rain marks in the results of JORDER and RESCAN. In the second image, there is a lot of rain and there is relatively large fog. It can be seen from the figure that a lot of rain marks remain in the results of JORDER, and many details in RESACAN are blurred. Relatively speaking, the method in this paper can still effectively remove the rain, and recover the rich texture background, which shows the effectiveness of the method in this paper in complex situations.

III. Summary and Outlook

In this paper, some experimental attempts have been made to remove rain marks from a single image. Although some progress has been made, there are still some shortcomings, which need to be further optimized in future research work.

Since objects may move in real life scenes, and ambient lighting and camera exposure parameters may change, it is nearly impossible to take a large number of photo pairs with and without rain in the same environment. Therefore, the rain removal algorithms based on deep learning are all trained on synthetic images. Most of the existing synthetic rain image datasets add rain marks of various shapes on the background image, but as the depth of the scene increases, the intensity of the rain marks decreases, and the occlusion effect of fog increases. If a more realistic rainmark dataset can be established, it will be of great help to improve the ability of the network to remove rainmarks.

The scale of existing deep neural network parameters is basically more than hundreds of thousands. The increase of a large number of network parameters significantly improves the network feature extraction ability and improves its task ability. However, the huge network structure limits its storage space and computing resources, Porting on the platform. In order to expand the application scenarios of deep learning network so that it can complete corresponding functions on a variety of small platforms, we should use model compression methods to optimize it. On the premise of ensuring the performance of rain removal, we should further reduce the network scale and improve its operating efficiency. It is the direction of our next research.

At present, the research ideas of model compression for deep learning networks are mainly divided into two categories: one is to compress the model through techniques such as quantization, pruning and knowledge distillation, and the other is to compact the original network by convolution operations. It can realize the functions of the original deep learning network with a small amount of calculation.

REFERENCES

[1] Yang F, Ren J, Lu Z, et al. Rain-component-aware capsule-GAN for single image de-raining [J]. Pattern Recognition: The Journal of the Pattern Recognition Society, 2022(123-).

[2] Wei Y, Zhang Z, Zhang H, et al. A Coarse-to-Fine Multi-stream Hybrid Deraining Network for Single Image Deraining [J]. IEEE, 2019.

[3] Porres I, Ahmad T, Rexha H , et al. Automatic exploratory performance testing using a discriminator neural network[C]// 2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW). IEEE, 2020.

[4] Yu C, Chang Y, Li Y, et al. Unsupervised Image Deraining: Optimization Model Driven Deep CNN[C]// MM '21: ACM Multimedia Conference. ACM, 2021.

[5] Lin X, Huang Q, Huang W, et al. Single Image Deraining via detail-guided Efficient Channel Attention Network [J]. Computers & Graphics, 2021, 97(2).

[6] Yang W, Tan R T, Feng J, et al. Deep Joint Rain Detection and Removal from a Single Image [J]. 2016.

[7] Sakaridis, Christos, Dai, et al. Semantic Foggy Scene Understanding with Synthetic Data [J]. INTERNATIONAL JOURNAL OF COMPUTER VISION, 2018, 126(9):973-992.

[8] Tang K, Yang J, Wang J. Investigating Haze-Relevant Features in a Learning Framework for Image Dehazing[C]// Computer Vision & Pattern Recognition. IEEE, 2014.

[9] Mao X, Li Q, Xie H, et al. Least Squares Generative Adversarial Networks [J]. 2016.

[10] Wang Z, Chen J, Hoi S. Deep Learning for Image Super-resolution: A Survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, PP (99):1-1.

Design and Research of Indoor Lighting Control System Based on the STM32

Yueyao Tian
Dept of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: 3247882993@qq.com

Lei Tian
School of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: tla02@126.com

Zhiyao Zhang
School of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: cokebjer@163.com

Wenxin Zhao
School of Humanities and foreign languages
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: wxh2324@126.com

***Abstract*—After the completion of the process of industrialization in the world, a large number of electrified equipment appeared, people's daily life is also more and more dependent on electricity security, electricity consumption has risen sharply. At present, various lights driven by electricity play a crucial role in our daily life, which is also the main cause of electricity consumption. We often notice that even when the room is empty, or during the day, the lights are still on, which is a waste of electricity. The waste of power resources caused by traditional lighting equipment cannot be ignored. Therefore, intelligent home furnishing equipment is the favored object of people. So here for a kind of indoor lighting control system is designed, in order to meet the lighting needs on the basis of more humane, more importantly, can save power resources. With the rapid application of automation in daily life and the continuous development of computer technology, lighting control has become increasingly intelligent and automated. The control function of the system is realized by the STM32F103 chip, the illumination is automatically detected by the BH1750 module, and the time parameters are recorded and fed back by the DS3231 clock module. The infrared reflection sensor is used to detect whether there are people in the room. When there is no one, the light will be automatically turned off to save energy. If there are people, when the ambient light intensity is lower than the set value, the light will be automatically turned on to ensure the comfort of the ambient brightness. And can use OLED module real-time display indoor light intensity, number of people and working time. This indoor lighting control system can not only automatically turn on the light according to the environment, but also detect whether there is a human body in the current environment and turn on or off the light by itself, which has more practical significance.**

Keywords-Lighting Control; STM32; Lighting Adjustment; Sensor; Intelligent

I. INTRODUCTION

Light consumption in daily life causes an unprecedented waste of electricity resources. Countries carry out research on this issue. Now, many products in the domestic market mainly use different step-down technologies to achieve energy saving, such as self-disaster transformers and magnetic saturation reactors, but these products have more or less problems. After the improvement, the equipment aging fast, eliminated products to the environment and other problems such as greater pollution.

This paper compares the control principle and main function structure of some indoor lighting control system, describes in detail the hardware structure of the indoor lighting control system and the design and implementation of the control system, and carries on various debugging [1].

II. THE OVERALL DESIGN

The traditional lighting control in China usually adopts manual control or timing control [2]. This method is time-consuming, inefficient, and will lead to a large degree of resource waste. The combination of camera monitoring and remote control greatly increases the cost of construction and will inevitably cause extravagance and waste of power resources [3].

Common domestic lighting control usually adopts some manual or timing control methods, such as the following common methods:

1: The lighting system is controlled manually, and the lighting device in a certain area is turned on or off by manual patrol, which is time-consuming and labor-intensive;

2: Use sound to control the lighting system, and use a sound sensor to judge whether the lighting device should be turned on. This method is inefficient, and all disturbances may cause the lighting device to work and cause a greater degree of resources waste;

3: Use optics to control the lighting system, mainly to control the lighting device according to the brightness in a certain area, this method is similar to method 2.

Although the traditional lighting device control methods have their own characteristics, they also have some shortcomings. Even some scenes with more lighting devices will be controlled by a combination of camera monitoring and remote control, which can indeed achieve better results, but this method greatly increases the cost of construction, and because this method is still manual Control, so it will inevitably still cause a lot of extravagance and waste of power resources [4].

In this paper, STM32F103 is used as the main control chip of the indoor light control system [4], and the design and implementation of the indoor light control system is carried out by combining DS3231, OLED, infrared detection and BH1750 light intensity detection modules. The BH1750 light intensity detection module detects the brightness of the current environment. DS3231 clock module is to realize the real-time display of time and working time. TCRT5000 infrared reflection detection module realizes the function of human body induction [5]. By conducting infrared reflection induction in a specific area, it automatically controls the opening or closing of indoor light control system [6]. OLED reminds users of the current state of indoor lighting control systems. The system framework is shown in Figure.1 below.

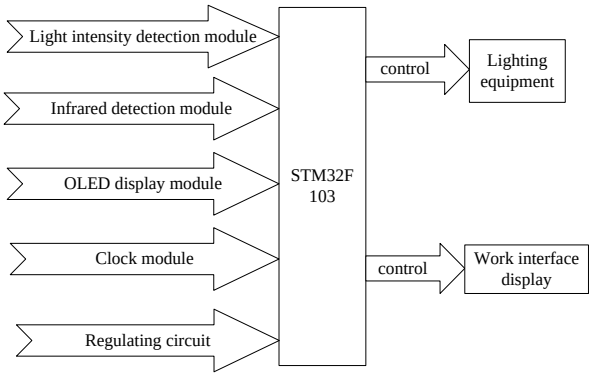


Figure 1. System block diagram

A. System control chip

In the indoor lighting control system, the position of the system control chip can be said to be very important [7], similar to the human brain, the instructions for the actions of each part of the body come from this. The control chip of the indoor lighting control system not only needs to control its own modules, but also needs to control some externally extended modules, so that the entire indoor lighting control system can operate [8].

Through the characteristics of 51 series, 32 series, 15 series and other types of single-chip microcomputers, combined with the requirements of the indoor lighting control system, the selected control chip has the characteristics of high speed, low power consumption and anti-interference. In the design of this article, the selected control chip It is a STM32F103 single-chip microcomputer chip. For this single-chip microcomputer, all the above-mentioned features are possessed, and it has more advantages. STM32F103 is widely used in many smart home control systems such as rice cookers, washing machines, refrigerators, and air conditioners, especially in the application of simple or single-function small micro-control systems is very common.

B. Environmental testing unit

The processing of sensors and brightness in the indoor lighting control system is also not negligible [9-10]. In the design of the system in this paper, TCRT5000 is used to detect the human body in the current environment; BH1750 detects the brightness in the current environment in real time.

For the TCRT5000 photoelectric sensor module, it is an infrared reflective photoelectric switch realized by the TCRT5000 infrared photoelectric sensor. For the sensor, it is composed of two parts, the first is an infrared photodiode, and the second is a high-sensitivity phototransistor [11]. The output control signal can be reshaped by Schmitt circuit, which is reliable, safe and reliable. The stability is very good. When in use, the built-in infrared emitting diode in the sensor will continuously emit infrared rays.

When the intensity of the emitted infrared rays is low or there is no emission at all, then the photosensitive transistor will always be in an off state. At this time, the output terminal corresponding to the module will be in a low level state, and at the same time, for the indicator diode, it will always be in a extinguished state; if the detected object is in the detection range, then the infrared will be Is emitted back and has a greater intensity at the same time, then the phototransistor will be in a saturated state [12]. At this time, the output terminal corresponding to the module is in a high-level state, and the indicator diode will be lit.

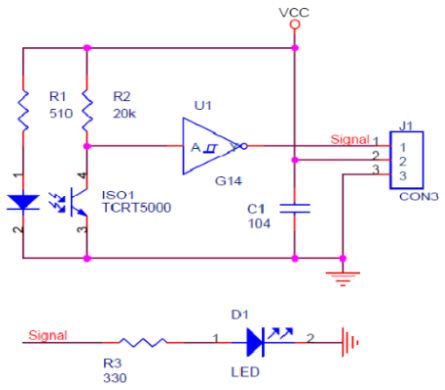


Figure 2. Infrared reflection module

According to the light intensity sensor model BH1750FVI, it is a digital light intensity sensor integrated chip in principle. The inside of the sensor is composed of four parts: ADC acquisition, photodiode, crystal oscillator and operational amplifier. For PD diodes, a photovoltaic effect is used to convert the input optical signal into an electrical signal, which is then amplified by an operational amplifier circuit, and then the voltage is collected by ADC, and then logic is used The circuit converts it into a 16-bit binary data, and then stores it in the internal register (it should be noted here that if the intensity of the light entering the light window is greater, the corresponding photocurrent will be greater.

At the same time, the voltage will be higher, so the voltage value can be used to judge the size of the light, but it should be noted that even if the voltage and the light intensity are corresponding, this relationship is not a proportional relationship, so the chip performs internally The corresponding linear processing, which also explains why not choose photodiodes, but choose integrated IC). At the same time, the sensor carries out the data line and the clock line. The microcontroller can use the I2C protocol to communicate with the sensor, can use the BH1750 working mode, and can also extract the illuminance data stored in the BH1750 register.

C. Clock module and display module

For the clock module model DS3231, it is an IPC real-time clock (RTC) in principle, with higher accuracy and lower cost, and an integrated temperature-compensated crystal oscillator (TCXO) is set inside. And crystals. For this device, there is a battery input terminal, after disconnecting the main power supply, it can still ensure that the timing is accurate enough. At the same time, through the use of an integrated crystal oscillator, the long-term accuracy of the device can be improved, and the number of components required in the production line can be reduced.

For, it can realize the provision of industrial grade and commercial grade temperature range, the number of pins used is 16 in total, and a 300mil so package is also used. RTC can maintain a variety of time parameters. If the number of days in the month does not exceed 31 days, the end of the month will be automatically adjusted and the leap year compensation function can also be realized.

The clock has two working formats, the first is 24 hours, and the second is 12 hours with AM\PM indication. Able to realize one programmable square wave and two programmable calendar alarm clock output. Both data and address can use the PC two-way bus to achieve a serial transmission, a voltage reference and a comparator that have completed temperature compensation to monitor the Vcc status, detect power failures, and provide reset output at the same time, In some necessary cases [13], it can realize automatic switching to the standby power supply. In addition, for the RST monitoring pin, it can also be used as a manual button to generate an external reset signal.

In the design of this article, the OLED display screen is chosen to design the display module [14]. This display screen has many names, and

sometimes it is also called an organic light-emitting semiconductor. This display screen was researched by a scholar in 1979. The characteristic of OLED display technology is that it can realize self-luminescence, has a wide viewing angle, fast reflection speed and low power consumption. The disadvantage is that the price is more expensive.

III. HARDWARE CIRCUIT DESIGN

The hardware circuit of the indoor lighting control system is mainly composed of STM32F103C8T6 MCU, DS3231 clock module, infrared detection module circuit, OLED display module and lamp tube circuit [15]. Combined with the requirements of indoor lighting control system control chip with high speed, low power consumption and anti-interference characteristics, this paper chooses the model of STM32F103 MCU chip.

The OLED display module and BH1750 light intensity sensor can be used together to obtain the brightness value of the current environment, so as to control the chip for the next processing judgment and issue corresponding instructions. In the clock module, users can adjust the indoor lighting control system with their own needs. Infrared detection module is set in the entrance and exit infrared reflection sensor, detection personnel in and out, so as to determine the number of people in the room. The lamp tube circuit is the main display part of the indoor lighting control system. The specific hardware circuit design is as follows.

A. Infrared Detection Module Circuit

Infrared detection module is composed of STM32 and CTRT5000 infrared detection module. After the indoor lighting control system is powered on, the CTRT5000 starts to work when it works normally. The infrared detection of the current environment is continuously carried out and the results are sent to the control chip. The control chip changes the state of the indoor lighting control system according to the returned results. The connection circuit between the infrared detection module and STM32 is shown in Figure 3.

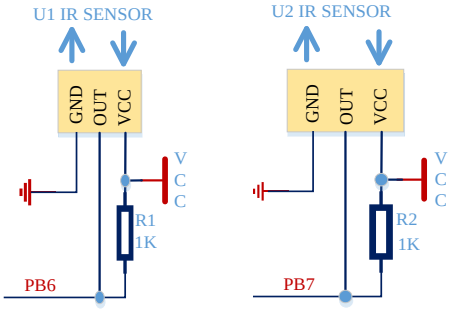


Figure 3. Connection circuit between IR sensor and STM32

B. Light Intensity Sensor Circuit

The light intensity sensor circuit of the indoor lighting control system is used to detect the light intensity. The BH1750 sensor is built with four components: an ADC acquisition, a photosensitive diode, a crystal oscillator, and an operational amplifier.

For PD diode, the use of light born v effect according to the input optical signals into electrical signals, then through operation amplifier circuit amplification processing, through the ADC to acquisition of voltage, after using logic circuit transformation, makes it a 16-bit binary data, and then stored in internal registers. This sensor carries on the data line and the clock line extraction, the micro-controller can use I2C protocol to realize the communication between the sensor, can use the BH1750 working mode, but also can extract the BH1750 register stored in the illuminance data. The BH1750 communicates in five steps: Step 1: Sends a power-on command. Step 2: Send the measurement command. Step 3: Wait for the measurement. Step 4: Read the data.

Step 5: Calculate the result. Finally, the light intensity data stored in the module is extracted by the master chip and displayed by the OLED.

C. OLED Display Circuit

The corresponding power supply input terminal of this module is VCC_IN, which supplies U1 and U4 circuits after depressurization by U2. A voltage of 7~7.5V is required to drive the OLED screen, which is provided by boosting the voltage through SSD1306's built-in charge pump. For SSD1306, it can support parallel, SPI and I2C serial protocols. This module is grounded by using BS0->BS2 pins, which have been set as SPI protocol and shared with U4 Chinese character library chip for SPI interface. During operation, the chip's corresponding chip pins are used to distinguish. During read and write operations, the CS pins corresponding to the chip will be set to a low-level state, which will be used as the corresponding display part of this project. OLED is connected to STM32 and BH1750 and DS3231. In addition to displaying normal working hours, it also displays brightness, number of people, current time, etc.

D. Rectifier Voltage Regulator Circuit

In the circuit of the lighting control system, AMS1117 series regulator has two versions, one is adjustable version, one is a variety of fixed voltage version, the role is to carry out 1A current output, while ensuring that the working pressure difference will not exceed 1V. The minimum differential pressure of the AMS1117 device is guaranteed to be less than or equal to 1.3V at the maximum output current, and to decrease as the load current decreases. SS34 is a patch Schottky diode used for instantaneous circuit rectification. The function of the rectifier circuit is to convert a kind of AC voltage in positive and negative changing state through a one-way conductivity of the diode, so that it becomes a one-way pulse voltage. Through the action of the AC power supply, the rectifier diode can achieve a periodic on or off state, by which the load can receive a pulse of direct current. Rectifier voltage regulator circuit is shown in Figure 4.

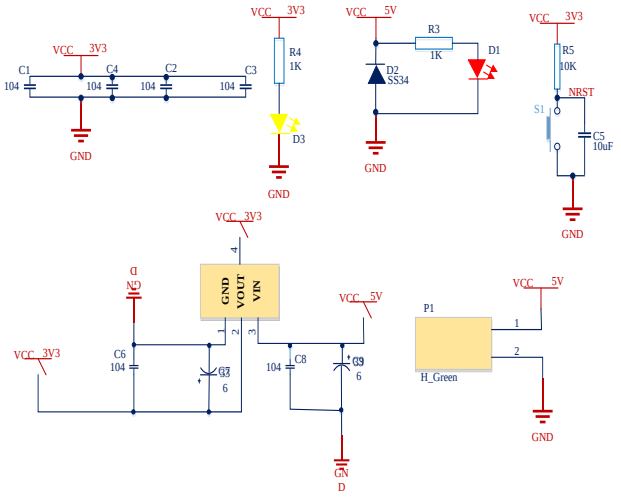


Figure 4. Rectifier voltage regulator circuit

IV. SOFTWARE PROGRAM DESIGN

A. System Software Design

In the research and development of any electronic product system, not only the design and implementation of hardware need to be emphasized, but the design and implementation of software programs are also equally important. At present, the development tools of various electronic systems can be described as a dazzling array of functions, but many senior developers still prefer to use KEIL for development, mainly because the pages are friendly and easy for developers to use since KEIL came out. Therefore, the software design part of the indoor lighting control system is also selected to be developed through KEIL and C language. Moreover, for the increasingly mature application and development of stm32 single-chip microcomputers, many basic codes and routines are also increasing, which makes it more convenient and simpler to develop

STM32 single-chip microcomputers based on KEIL in C language.

The research and development of the software part of the indoor lighting control system mainly adopts a process from a single sub-module to an overall realization, which facilitates the timely detection of errors and reduces the time for follow-up inspections:

1) The design of the environmental brightness detection module: The BH1750 is mainly used to detect the current environmental brightness in real time and transmit the data to the program of the main control chip.

2) OLED display module design: Mainly connected with the main control chip, it can realize the real-time display of the feedback data on the screen, so that the user can understand the current information program.

3) Clock module design: DS3231 performs accurate time calculation, and feeds back the program of real-time time and current working time.

4) Human body detection module design: It is mainly a program that detects the human body in the current environment and makes corresponding operation feedback.

Each subroutine relationship is shown as Figure 5.

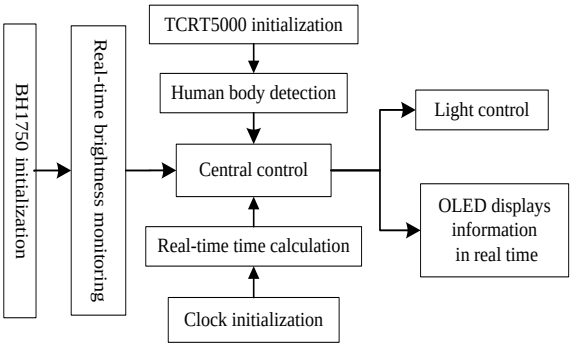


Figure 5. Relationship among subroutines

B. Ambient Brightness Detection

Intelligent lighting detection of the current environment brightness subroutine mainly uses BH1750 to detect the current environment brightness. The BH1750 light detection sensor needs to be initialized and the environment brightness is detected according to the instructions of the control chip. The environmental brightness detection subroutine flow is shown in Figure 6.

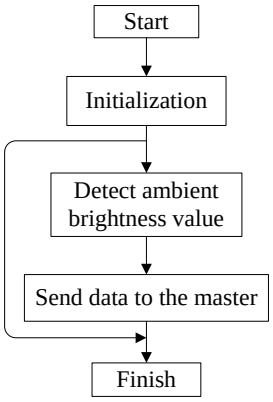


Figure 6. Brightness detection subroutine

When detecting the brightness of the current environment, this paper adopts the method of twice detecting, that is, continuously detecting the brightness of the current environment twice. When the two detection results are consistent, the collected ambient brightness value will be sent to the control chip; otherwise, "0" will be sent.

C. OLED Display Programming

The display module program of this design needs to be set for the IO connected by the MCU and the OLED module, and the IO port is set as output. Initialize the OLED module. Use functions to display characters and numbers on the OLED module.

D. Clock Subroutine Design

The design of the clock module program through STM32 analog I2C signal channel, then configure IO pin, and initialization function, in the main function to call the time data of the module.

E. Body Detection Subroutine

Human body detection subroutine is a characteristic function of intelligent lamp, which can intelligently detect whether there is human body in the current environment. The control chip will intelligently control the opening or closing operation of the intelligent lamp according to the detection results. The feature is very convenient and friendly, especially at night. The human body detection subroutine flow is shown in Figure 7.

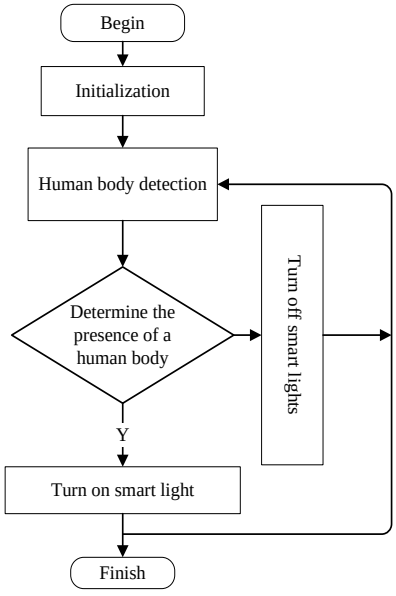


Figure 7. Human body detection subroutine diagram

Once powered on, the smart lamp circulates through the current environment. The intelligent light will be turned on automatically when someone is detected, and will be turned off when no human body is detected.

F. Main Programming

The main program of indoor lighting control system is to initialize and circularly call subroutines such as clock module, display module, human body detection module and light intensity detection module. In this way, the system can be ensured to run smoothly. The main program flow is shown in Figure 8.

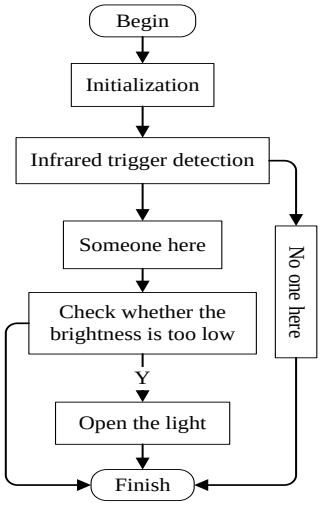


Figure 8. Main program flow

After the indoor lighting control system is powered on the initialization settings, mainly including the initialization of some sub-modules and the declaration of common functions and so on. Make circular calls to sub-modules and between sub-modules.

This chapter mainly uses KEIL and C language to research and develop the software part of the indoor lighting control system. In the process of designing and implementing subroutines such as clock module, display module, human body detection module, light intensity detection module, etc., not only the theoretical knowledge learned is consolidated, but also his practical ability has been strengthened.

V. FUNCTION IMPLEMENTATION AND TESTING

The selection of main modules and components of the indoor lighting control system and the design of software and hardware have been implemented in the previous chapters. The following is mainly to integrate and debug the completed indoor lighting control system

The indoor lighting control system is composed of STM32F103, human body detection module circuit (infrared detection), clock circuit, display

circuit, light intensity detection circuit and lamp circuit.

The first thing to do during and after the realization of the electronic system is debugging. Debugging during the implementation process can help to find and correct the shortcomings in the implementation process as early as possible; and the debugging after the implementation can help to find and correct the shortcomings of the system in time. There are many ways to realize the debugging of electronic systems, including software and hardware debugging and module debugging, system joint debugging and simulation debugging.

This system adopts the method of single module debugging and simulation debugging. The indoor lighting control system is debugged through hardware debugging and software debugging respectively. Hardware debugging mainly uses multimeter, oscilloscope and other tools to check the hardware circuit of each submodule; software debugging mainly uses KEIL development tools to detect human body detection module subroutines, button module subroutines, various working mode subroutines, etc., as well as the entire indoor lighting The program of the control system is debugged and a file that can be recognized by the control chip is generated.

The whole system is shown in Figure 9.

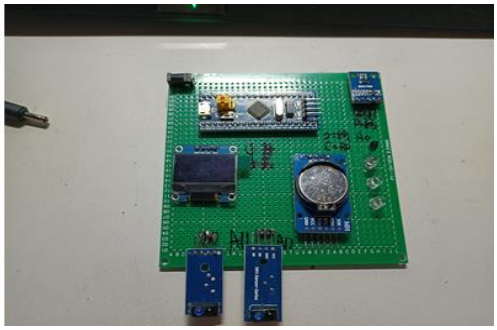


Figure 9. The photo of the system

After the power supply is connected, the system enters the standby state. The screen shows the current working status and ambient brightness with the number of detected people.

If the infrared detection module works properly and detects human body entering the room, the BH1750 module starts to work. If the ambient light is detected below the set value (500 lumens), the system will automatically turn on the light for brightness enhancement. If the infrared detection module detects a human body, but the ambient light is greater than the set value (500 lumens), the system will not turn on the light. The two cases are shown in Figure 10. and Figure 11 respectively.

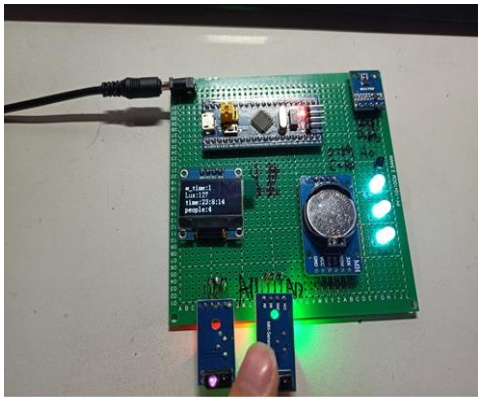


Figure 10. The ambient light is below the set value

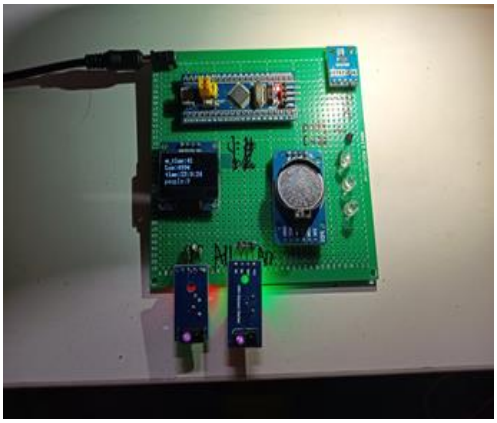


Figure 11. The ambient light intensity is too high

VI. CONCLUSION

In order to achieve green indoor lighting, this paper uses STM32F103 chip to realize the automatic control system of illuminance. The light is automatically detected by the BH1750 module, and the feedback is recorded by the DS3231 clock module. Use infrared sensors to detect whether there are people in the room. When no one is around, the lights automatically turn off to save energy. If there is someone, when the ambient light intensity is lower than the set value, the light will automatically turn on to ensure the comfort of the ambient brightness. This indoor lighting control system can not only automatically turn on the lights according to the environment, but also detect the number of people in the current environment to turn on the number of lights, which is very practical and has high promotion value.

ACKNOWLEDGMENT

This work was partly supported by the undergraduate innovation and entrepreneurship plan (No.202211664099X, 202211664012 and 202211664021).

REFERENCES

[1] Bhartiya, G., and P. Pathak. "Intelligent Lighting Control and Energy Management System." 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) Feb. 2020, pp.86-89, doi:0.1109/PARC49193. 2020.236563.

[2] L., Heng, et al. "Development of an Indoor Intelligent Lighting System Based on Occupants Positioning." International Journal of Pattern Recognition and Artificial Intelligence, 2021. doi:10.1142/S0218001421590412.

[3] As, A, et al. "Efficient photodetector placement for daylight-responsive smart indoor lighting control systems." Journal of Building Engineering, vol.42, Oct. 2021, pp.1-12, doi:10.1016/j.jobe. 2021.103013.

[4] E. Shao, and X. C. Peng. "Research on Light Distribution Optimization Design of Road Lighting LED Lamps." Western China Communications Science & Technology, vol.3, Feb.. 2019, pp.148-152. doi:10.13282/j.cnki.wccst. 2019.03.043.

[5] N. A. Binti Mohd Arifin and N. M. Thamrin, "Development of Automated Microcontroller-Based Lighting Control System for Indoor Room Implementation," 2018 4th International Conference on Electrical, Electronics and System Engineering (ICEESE), 2018, pp. 82-86, doi: 10.1109/ICEESE.2018.8703500.

[6] W. Byun, Y. Jin, Y. Kim and J. Lim, "Design of Lighting Control System Considering Lighting Uniformity and Discomfort Glare for Indoor Space," 2018 International Conference on Platform Technology and Service (PlatCon), 2018, pp. 1-6, doi: 10.1109/PlatCon. 2018.8472750.

[7] Seyedolhosseini, N. Masoumi, M. Modarressi and N. Karimian, "Illumination Control of Smart Indoor Lighting Systems Consists of Multiple Zones," 2018 Smart Grid Conference (SGC), 2018, pp. 1-4, doi: 10.1109/SGC. 2018.8777883.

[8] F.Y. Deng, Y.W. Qiang, Y.Q. Liu, et al. "Adaptive Parametric Dictionary Design of Sparse Representation Based on Fault Impulse Matching for Rotating Machinery Weak Fault Detection." Measurement Science and Technology , 2020, vol. 31, pp. 117-120. doi: 10.1088/1361-6501/ab6f2f.

[9] T., Tsagkatakis, P. P., Smirnakis and Tsakalides. "Adversarial Dictionary Learning for a Robust Analysis and Modelling of Spontaneous Neuronal Activity." Neurocomputing, 2020, vol. 38, pp. 188-201. doi: 10.1016/j.neucom. 2020.01.041

[10] L. B. Wang, C., Chen, Y. J. Li. "Dictionary Learning Algorithm Based on Weighted Least Square." Systems Engineering and Electronics, vol. 33, 2011, pp. 1896-1900. doi:10.3969/j. Issn.1001-506X. 2011.08.41.

[11] Naderahmadian, Yashar, Beheshti, Soosan, Tinati and Mohammad: Correlation Based Online Dictionary Learning Algorithm. IEEE Transactions on Signal Processing, vol. 64, 2015, pp. 592-602, doi:10.1109/TSP.2015.2486743.

[12] U. I. Kizgut and Ersin. "Lossy Compressive Sensing Based on Online Dcitionary Learning." Computing and informatics, vol. 38, 2019, pp. 151-172. doi: 10.31577/CAI_2019_1_151.

[13] N. Takayuki, B., Yukihiro and Kiya, Hitoshi. "Secure Overcomplete Dictionary Learning for Sparse Representation." IEICE Transactions on Information and Systems, vol. E103D, 2020, pp. 50-58. doi: 10.1587/transinf. 2019MUP0009.

[14] Li Yun, Fu Ai-Ping. "Designing Metadata for Chinese Dictionary Entries." Data Science Journal, vol. 6, 2007, pp. 853-856, doi: 10.2481/dsj.6.S853

[15] T. Huu and Monga Vishal. "Fast Low-Rank Shared Dictionary Learning for Image Classification." IEEE Transactions on Image Processing, vol. 26, 2017, pp. 5160-5175, doi: 10.1109/TIP. 2017.2729885.

Research on Oil Well Data Cleaning System

Yao Feng
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 2352933419@qq.com

Li Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com

***Abstract*—In the information age, with the continuous development of Internet technology, information data occupies every field of contemporary society. The development of the big data age makes these data more and more prominent. While users read the information they need from these massive data, data quality has also become a concern of users. A large number of data are preprocessed before data analysis, such as some duplicate values, missing values deal with inaccurate and other abnormal data, and filter the data through the data cleaning system to improve the standardization of the data, so as to improve the analysis efficiency of the data, reduce some unnecessary expenses, and save time and effort. The data cleaning system in this paper is implemented based on flash framework. Taking Python as the main language for data cleaning, technical cleaning and standard integration are carried out for some structural problems, duplication problems and missing problems of some different source data. Through the processing of abnormal data, the data quality and data analysis efficiency are greatly improved.**

Keywords-Data Cleaning; Cleaning System; Oil Well; Outlier Handling

I. INTRODUCTION

Data is recorded with identifiable symbols or numbers to reflect objective things, such as pictures, numbers, words, etc. in the oil production industry, the data value of each attribute reflects various indicators of oil wells, and each oil well has a certain range. Once it exceeds this range, it will affect the subsequent oil exploitation. In order to realize the normal exploitation of oil wells and realize the maximum oil production, data quality becomes very important. Therefore, users should clean some data affecting data quality from massive data, so as to improve the quality of data and make the oil exploitation work go smoothly. on this basis, this paper studies the processing of invalid data in oil wells. In order to optimize the data and facilitate the subsequent analysis, this paper proposes and designs an oil well data cleaning system based on Web.

II. RESEARCH BACKGROUND AND CURRENT SITUATION

In contemporary society, oil, as the backbone of modern industry, is of great significance to national development and is closely related to people's livelihood. Although China's oil and gas resources are relatively rich compared with other countries, compared with the world average, the per capital percentage of oil in China is still relatively low. Therefore, China must take effective measures to increase oil production and formulate reasonable plans to improve China's economy. Increase oil production and realize the maximum utilization of oil resources. Therefore,

the research on data cleaning system is very necessary for industrial production.

There are many researches on data cleaning at home and abroad [1], including: combining clustering algorithm, association rules and wavelet neural network to check and modify the abnormal data in the sensors and equipment that can be cleaned in the monitoring data. At the same time this method has a disadvantage that it needs to carry out sencon-ary fitting correlation model for the data in the sliding window, [2] which not quickly clean the data that needs to be cleaned, In addition, the monitored data is regarded as the time series of each state quantity, and the iterative method is used to correct the abnormal data in the data. [3] This method also has an obvious disadvantage that there is a large deviation between the original value and the modified value, which causes some damage to the integrity of the data. There are also methods that can deal with outliers by constructing boundary models, but the processing results are inaccurate, which damages the continuity of data. There are also methods that use gravity search algorithm to identify errors and repair data [4], but the real-time performance of data is not very accurate. Therefore, the existing data cleaning methods only deal with the abnormalities of local state quantities, the processing of attribute correlation between data is not very good, and there is some damage to the continuity and integrity of data. Therefore, in order to reasonably clean the data, we should deeply express the data characteristics [5].

In recent years, with the rapid development of economy and the substantial improvement of information technology, many enterprises also have many problems in the design of data cleaning system, and there are still many aspects to be improved in the production of cleaning system [6]. Therefore, the manufacture of cleaning system has a great application prospect for industrial production. It can depend on two factors:

- Compared with the original manual processing stage, the system can effectively process the abnormal data without spending a lot of time, manpower and financial resources to process these abnormal data, greatly improve the data quality and maximize the utilization of resources, which can not only save costs but also improve the utilization of resources.
- It has an immeasurable impact on China's economic development. The effective processing of data can obtain more accurate analysis results. According to these effective data, we can make appropriate adjustments to the mining scheme, improve oil production and be conducive to the country's economic development.

Based on the above two factors, we can understand that there is a lot of room to improve the development of oil well data cleaning system based on Web. The system is not only conducive to the effective utilization of resources, but also can save cost, time and labor, which has a great impact on national development and industrial production [7].

III. Demand Analysis of Oil Depot Data Cleaning Method Based on Web

As the starting point of system development, requirement analysis has guiding significance for system development, so requirement analysis plays a very important role. It is to make a comprehensive analysis of the problems to be solved, and then know what problems need to be solved. The person who makes the system must first understand what customers need, and then realize these in the developed system [8]. Demand analysis is the link between developers and

customers. Only by understanding the needs of users can we make a system to meet the needs of customers. Next, I will analyze the functional and nonfunctional requirements.

A. Functional Requirements Analysis

The system mainly provides a data cleaning platform for industrial production departments to clean oil well data. The functions of the system mainly include:

- Missing value processing

In the oil well data, there are many missing values due to mechanical and human reasons [9]. Mechanical factors are the failure caused by mechanical data collection or problems in saving data, such as memory damage, and human factors, such as arbitrary data modification and input errors. The existence of these values will have a great impact on the subsequent analysis results. Therefore, it is necessary to deal with these missing values, to improve data quality [10].

- Duplicate value processing

In the data set, there will be many duplicate values due to various errors. Some attributes can only appear once, so a large number of duplicate values will have an immeasurable impact on the data quality. Therefore, there is a great demand for the processing of duplicate values.

- Outlier handling

In the data, due to collection errors or system problems, some values that do not conform to common sense or values that do not conform to the defined data type may appear in the data. For example, abnormal data such as text and multiple decimals may appear in the data. In the process of cleaning [11], these abnormal values cause errors in the whole analysis and affect the whole progress of the analysis. Therefore, we must pay attention to the handling of outliers [12].

B. Nonfunctional Requirements Analysis

The evaluation of image quality can be divided According to the feasibility analysis, we can determine whether the production personnel will be satisfied with the cleaning system, according to the feasibility analysis of three aspects: technology, economy and operation. The data cleaning system is helpful for industrial production personnel to filter messy data, improve data quality and facilitate subsequent analysis [13].The following is a brief analysis of three specific aspects.

Technical feasibility: the system adopts the springboot framework and the main language is python. The system adopts the Android system architecture [14], and uses JSP technology to store information and standardize functional modules. Teachers and students can operate on mobile phones. According to the current mobile phone configuration, its own system performance is very high and it is very easy to realize.

Economic feasibility: the early data cleaning is based on manual operation. Under such operation, not only the data processing results are not ideal, but also a lot of time and energy will be spent on it. Before designing a data cleaning system, it is necessary for designers to think about its cost and the amount of resources. Catering to the needs of industrial producers, there will be objective economic benefits after development.

Operational feasibility: the system only needs to run on the computer, and then use the database to import the data to be cleaned and operate to process the data. The use process is not complex, the operation is simple, and it does not need to spend too much human and financial resources. It does not need professional skills to operate the system and process the data. It has strong interactive performance and good research significance, It is of great significance to industrial production [16].

IV. Design of Oil Well Data Cleaning System

A. Cleaning Principle

In deep learning, the quality of data sets In the final analysis, data cleaning is to conduct in-depth analysis on the main causes of dirty data, find out the form in which these dirty data exist in a large number of data to affect the analysis results, use the current existing technology and optimize its processing to find dirty data, and change these detected data into values that meet normal needs after processing. The idea of data cleaning mainly uses the backtracking method [17], which takes the detected data as the starting point, makes a systematic analysis of each detail of the data flow direction, and summarizes a widely used data cleaning algorithm and related rules, which can be applied to various types of data cleaning systems [18].

B. Web Implementation

The web was mainly used for browsing static interfaces in the early days. These static interfaces are written in HTML and placed on the server. Users use the browser to request the web page on the server through the HTTP protoco. After the [19] web server software obtains the request sent by the user, it reads the identified resources [15]. When the client receives the browser sent by the message header, the browser parses the HTML data of the response, Show different vivid HTML pages to the client. However, with the continuous development of the network, more and more businesses begin to develop online, so the web application based on the network has become not simple. The resources accessed by users are not only the static pages stored on the server hard disk. A large number of applications require the web page information dynamically generated based on the user's request. It is even more difficult to extract some relevant information from the database and return a generated page to the customer through some operations.

How to achieve:

It is realized through the server software of HTTP protocol, which reserves the interface that can be extended in advance [20]. What the user needs to do is to provide relevant expandable functions according to the rules. When the client request is transmitted to the web server software, it will judge whether it is to visit the relevant extended functions provided by the user. If it is judged to be yes, the relevant programs written by the user will be processed. After the processing is completed, the program will return the processing results to the client.

C. System Framework

Flask is a web framework of python, which is very popular at present. The flexibility of Python language also makes flask have the same characteristics. Therefore, the biggest advantage of flask is that it is simple and lightweight, and developers can flexibly use the features to be developed, Let developers freely and flexibly compatible with the features to be developed. Whether it is user portraits or product recommendations, python has a very heavyweight advantage. Because flash is relatively light and simple, it is easy to start, and the cost of trial and error is low. Flask's framework also brings a more image extraction mechanism, which can work on almost any platform of Python. Flask's data access framework can also solve some difficulties that occur frequently when using the database [21].

D. Function Specific Design

- Pretreatment stage

Preprocessing is divided into two steps: one is to adjust the data to an uploadable file. Second, look at the data. Through manual viewing, you can have a more intuitive understanding of the data,

find some obvious errors at the beginning, and make some simple preparations for later processing.

- Missing value processing

Missing value is the most common problem in data. There are many filling methods for missing value: one is to fill in the missing value according to their own knowledge or experience; the other is to fill in the missing value according to the calculation results of the same indicators, such as filling in the missing value according to the calculation results of different indicators; the third is to fill in the missing value according to the calculation results of different indicators. The mean filling method is adopted in this design. The calculation results of the same index are used for summation, and then the average value is taken to fill in the missing data. This method has the advantages of simple calculation, fast filling speed and close to the real value [22].

- Duplicate value processing

In some attributes, some data can only appear once, and sometimes errors will occur, resulting in data duplication. Therefore, the processing of duplicate values should be realized.

For the processing of duplicate values, first find the duplicate values in the data set and save them in one list, save the duplicate index to another list, and then take out the average value of the current attribute and take out the index value for replacement.

- Outliers

Some unreasonable values will appear in the data, such as text classes or abnormal values beyond the normal range in the data set.

For values beyond the range, we will check and clear them according to the previously defined range, and then use the filling method. For text and decimal types, we can clear them by defining data types, and finally use the missing value filling method.

E. Flow chart system function

According to the brief description of specific functions, the relevant functional flow chart of oil well data cleaning system can be obtained. The functional flow of the system is shown in Figure 1.

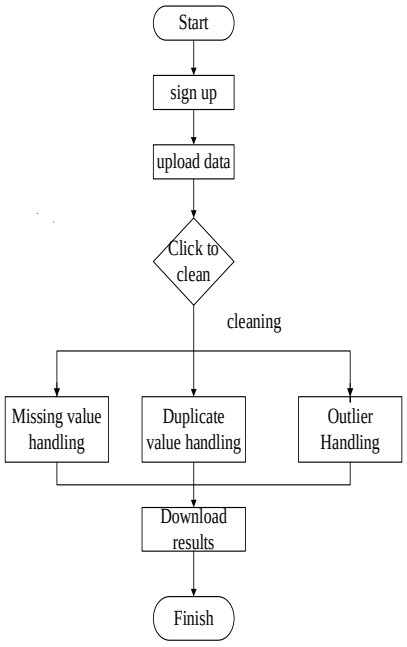


Figure 1. System function flow

F. Database implementation

MySQL is a relational database, which can be used in a variety of storage because it has multiple storage engines. It has many advantages: high speed, small volume, easy installation, etc. Most developers use it as the primary database. Now, under the name of Oracle, it is a kind of well used and free database, which can be used in small and medium-sized databases. If it is designed to be easy to use and support all kinds of locks, it can be used well in concurrent processing, and the types of MySQL fields meet the needs of most designers. The addition of different types of transactions also

makes the storage of data more favorable and improves the accuracy. It is not easy to cause repeated write and other problems. The management of permissions makes the distinction between superiors and subordinates more perfect, and it becomes easier to set permissions, so that the security of data is more and more guaranteed. Based on the above excellent points, I take MySQL database as the main database I develop and use. Its driver management is shown in Figure 2:

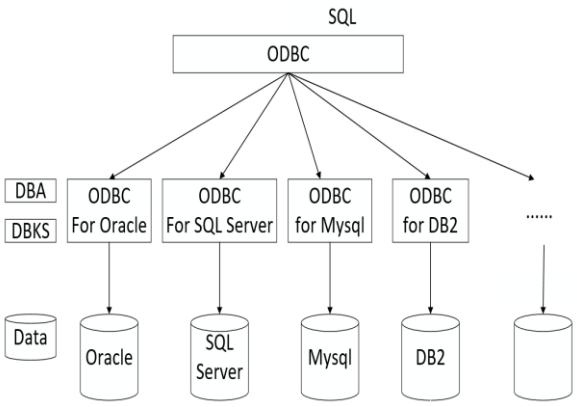


Figure 2. Database driver management diagram

V. Data Cleaning System of Oil Depot Based on Web

A. Function description

The web-based oil well data cleaning system is divided into two parts: the front-end interface display and the background data processing system, which are implemented based on the flash framework. The specific functions are as follows:

User: the user registers and logs in. After the user logs in successfully, the page jumps to the home page [25]. The user uploads the relevant oil well data to be cleaned according to the needs. After parsing the excel file, the data cleaning system cleans the data in the file. After cleaning, a link will be returned. You can apply to download the cleaned results by pasting it into the browser. Specific cleaning rules need to be defined by ourselves.

Its specific functions are as follows:

- User login registration

When users log in the system for the first time, they need to register to determine the login account. Users need to fill in their own user name and set the login password. When logging in directly, the system will automatically prompt whether there is a login account and whether the password is wrong. If there is an error, the system will enter it for the second time.

- User information management

Users can modify their own information in my home page. The information filled in when logging in for the first time is saved on my home page.

- Data cleaning

The user uploads the data to be cleaned, and then the system automatically parses it. After parsing, the data is processed according to the set range, mainly dealing with some missing values, duplicate values and abnormal values for embedding function, and can also use the tag of JSP to parse and convert it into HTML tag, or use XML tag.

- Download Results

After that, you can choose to download and save the cleaned data locally, and users can conduct in-depth analysis. Facilitate the next step of work.

B. Cleaning data upload

After data cleaning, a link will be returned. Copy the link to the browser to view the cleaned results, download and save the results. As shown in Figure 3.

Figure 3. Data import

C. Cleaning result output

The collected data is shown in Figure 4.The results after cleaning are shown in Figure 5.According to the comparison between figure 4 and figure 5, we can see that the red box indicates that the system fills in the missing value, the green box indicates that the missing value is filled in after the duplicate value is also deleted, and the yellow box indicates that the abnormal value is deleted and filled in. The system has basically realized the cleaning of messy data.

2	35-55	3.5			2.5		11.67	73.6	3		
3	54-61	5.3	11	229.75		5.4	13.97	51.6	4	1.9454	
4	21-51	3.4	32.93		2.42		8.37	59.88	3		
5	56-46	2.8					6.96	57.7	2		2.82
6	31-49	3.9	38.17	244.83	2.49	3.24	11.75	71.13	3		
7	32-65	3.1	22.37	262.56	2.36	3.34	15.23	70.35	2	15.6	2.4
8	32-65	7.3	15.02	239.97	2.44	3.75	10.79	53.8	2	25.1	3.5
9	29-54	2.8	46.94	235.14	2.51	3.61	9.66	61.4	2		
10	29-54	6	15.6	251.31	2.45	2.78	12.19	50.8	3		
11	34-28	5.5	48.42	210.75	2.55	6.8	5.65	55.8	3	18.0+0.5	2.2*2.3
12	34-57	5.4	11.74	243.55	2.52	12.24	11.5	46.05	3	30.0+0.6	2.4-2.5
13	34-54	2.8	26.54	239.97	2.47	8.4	10.5	60.3	3	25.0+0.51	2.04
14	34-54	4.1	19.36	232.11	2.51	12.04	9.2	51.41			
15	13-67	3.8	17.9	231.6	2.55	9.9	9.1	52.9	3	14.5	1.7
16	13-67	1.6	34.8	224.4	2.52	5.4	7.9	60.1			
17	05-105	4	38.46	229.75	2.5	5.8	8.8	62.45	2	26.7	2.82
18	05-105	3.9	18.23	251.23	2.43	7.1	12.37	57.67	3		
19	32-23	1.4	17.7	238.3	2.51	4.2	10.47	49.3	0.2	15	3.5
20	13-65	5.3	16.06	243.13	2.49	9.47	11.03	52.4	3	25.6	1.95
21	34-19	6.9	31.57	213.57	2.62	17.59	5.47	13.3	0.2	23.8	3

Figure 4. Data collection

2	35-55	3.5	31.37	234.57	2.5	6.91	11.67	54.38	3	17.65	3.1
3	54-61	5.3	31.07	229.75	3.06	5.4	13.97	51.6	4	1.9454	3.52
4	21-51	3.4	32.93	234.23	2.42	7.23	8.37	59.88	3	17.29	3.43
5	56-46	2.8	30.93	234.49	3.22	6.64	10.89	57.7	2	16.89	2.82
6	31-49	3.9	38.17	244.83	2.49	3.24	11.75	54.6	3	17.64	3.63
7	32-65	3.1	22.37	234.11	2.36	3.34	11.47	54.14	2	15.6	2.4
8	32-65	7.3	31	239.97	2.44	3.75	10.79	53.8	2	25.1	3.5
9	29-54	2.8	46.94	235.14	3.25	3.61	9.66	61.4	2	17.1	3.15
10	29-54	6	31.45	234.8	2.45	2.78	12.19	50.8	3	17.1	3.71
11	34-28	5.5	48.42	210.75	3.36	6.8	11.11	55.8	3	17.48	2.96
12	34-57	5.4	31.48	243.55	3.01	12.24	11.5	46.05	3	17.84	3.47
13	34-54	2.8	26.54	239.97	2.47	8.4	10.5	60.3	3	17	2.04
14	34-54	4.1	31.77	232.11	3.07	12.04	9.2	51.41	4.19	16.99	3.61
15	13-67	3.8	31.1	231.6	2.78	9.9	9.1	52.9	3	14.5	1.7
16	13-67	5.39	34.8	224.4	3.39	5.4	10.84	60.1	4	17.6	3.04
17	05-105	4	38.46	229.75	2.5	5.8	8.8	62.45	2	26.7	2.82
18	05-105	3.9	31.86	234.58	2.43	7.1	12.37	57.67	3	17.37	2.87
19	32-23	5.35	31.73	238.3	3.39	4.2	10.47	49.3	4.16	15	3.5
20	13-65	5.3	31.45	243.13	2.49	9.47	11.03	52.4	3	25.6	1.95
21	34-19	6.9	31.57	213.57	2.84	6.98	11.55	13.3	3.57	23.8	3

Figure 5. Cleaned data

D. Functional testing

The results are shown in Table 1. According to the above tests, we can know that the pre designed functions of the system can be basically realized smoothly without other errors during operation.

E. Operating environment

JSP is a popular service site at present. You can view it when you need JSP interface and input the website you want to place them. JSP is a back-end language. It is developed based on Java code. It shows a dynamic effect by inserting the data obtained from the operation of Java code in the file into the static page. It can use the tag of XML file Using this configuration can reduce the dependence of the code on the server platform and run everywhere, which is also due to the Java virtual machine. JSP compiler can convert the code into the original binary code, and then run the binary code directly, which is one of the advantages of JSP. Compared with pure servlets, JSP can easily write or modify HTML web pages without facing a large number of print statements. It reduces the burden of development, so I use JSP as my page language.

TABLE I. TEST CONTENTS AND RESULTS

Test content	Expected results	Actual results	judge
Enter user name and password	Login successful	Same as expected	adopt
Upload correct data sheet	Upload successful	Same as expected	adopt
Upload error data table	report errors	Same as expected	adopt
Are duplicate values processed	Processed	Same as expected	adopt
Are missing values populated	Filled	Same as expected	adopt
Are outliers handled	Processed	Same as expected	adopt
Whether data can be saved	Successfully saved	Same as expected	adopt

VI. SUMMARY

The oil well data cleaning system based on Web has great potential and development value. The implementation based on web in this paper integrates the design mode, further improves the scalability and maintainability of the code, improves the concurrency of the system, and improves the concurrency ability of the system. When a large amount of data needs to be cleaned, it can ensure the normal operation of the system. Data mining technology is widely used in medical treatment, energy, retail, automobile, finance and many other fields to provide decision-making and suggestions by mining valuable information. Medical data mining can provide targeted guidance to patients, predict the change and development trend of physical health status, and take preventive measures. Accurate analysis can reduce over treatment and insufficient treatment; Use big data to analyze energy purchases, so as to predict energy consumption, improve energy efficiency and reduce costs by managing energy users; For retail enterprises, data mining technology can well integrate all kinds of information, help enterprises master customer needs, and realize precision marketing and personalized services; With the help of data mining technology, insurance companies ncan fully understand drivers' driving habits ad driving behavior, provide different types of insurance products, and so on. The widespread existence of dirty data leads to the limited available data in the process of data mining. Data cleaning is particularly important. Different data cleaning methods have their own advantages and disadvantages, and the research on data cleaning methods will be more in-depth.

REFERENCES

[1] Peng Nan Research and implementation of cleaning rule mining system for dirty data [D] Shanghai: Donghua University, 2018.

[2] Hu daiguo Design and implementation of data cleaning scheme for mobile comprehensive information system [D] Shandong: Shandong Normal University, 2018.

[3] Qu Wanrong Urban road traffic flow data cleaning technology and system implementation [D] Zhejiang: Zhejiang University of technology, 2017.

[4] Wang Jiang Research on data cleaning technology and design and implementation of cleaning framework [D] Inner Mongolia: Inner Mongolia University, 2016.

[5] Wang Kui Research and application of Hadoop based traffic data cleaning [D] Guangdong: Guangdong University of technology, 2017.

[6] Sun Chang Research on SCADA data cleaning method of wind turbine [D] North China Electric Power

University; North China Electric Power University (Beijing), 2018.

[7] Li Ningning Research and implementation of optimization technology in big data cleaning system [D] Heilongjiang: Harbin Institute of technology, 2016.

[8] Wang Liping Research on on-line cleaning technology of heat exchanger based on scaling and scale dissolution mechanism [D] Beijing: Beijing University of chemical technology, 2017.

[9] Zhang Xiaoli Research on data cleaning framework and missing value reconstruction method [D] Shaanxi: Xi'an University of technology, 2015.

[10] Fang Xiaoli, Liu Xia Research on data cleaning method for document visual analysis [J] Journal of University Library and information science, 2021,39 (06): 56-60.

[11] Yuan Yanwei, Xu Ling, Ji Fuhua, Guo Dafang, an SA, Niu Kang Big data cleaning method and experimental optimization of agricultural machinery operation [J] Journal of agricultural machinery, 2021,52 (06): 35-42.

[12] Xu Sijia, Wang Xiang, Zhao Chenglin, Xu Fangmin Research on data cleaning methods of industrial machinery and equipment [C] / /. Proceedings of 2020 China information and communication conference (CICC 20202020:347-352.)

[13] Data Cleaning in Cloud Platform [J]. International Journal of Recent Technology and Engineering, 2020, 9(1).

[14] Jinlin Wang, Xing Wang, Yuchen Yang, Hongli Zhang, Binxing Fang. A Review of Data Cleaning Methods for Web Information System [J]. CMC: Techniques Computers, Materials & Continua, 2020, 62(3).

[15] Yogita Bansal, Ankita Chopra. Data Cleaning for Large Data Sets [J]. International Journal of Recent Technology and Engineering (IJRTE), 2020, 8(6).

[16] Yanli Bai. Data cleansing method of talent management data in wireless sensor network based on data mining technology [J]. EURASIP Journal on Wireless Communications and Networking, 2019, 2019(1).

[17] Candelotto Laura, Grethen Klara J., Montalcini Camille M., Toscano Michael J., Gómez Yamenah. Tracking performance in poultry is affected by data cleaning method and housing system [J]. Applied Animal Behaviour Science, 2022(prepublish).

[18] Liu Shengjie, Li Guangye, Jiang Shize, Wu Xiaolong, Hu Jie, Zhang Dingguo, Chen Liang. Investigating Data Cleaning Methods to Improve Performance of Brain–Computer Interfaces Based on Stereo Electroencephalography [J]. Frontiersin Neuroscience, 2021, 15.

[19] Shi Xi, Prins Charlotte, Van Pottelbergh Gijs, Mamouris Pavlos, Vaes Bert, De Moor Bart. An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge [J]. BMC Medical Informatics and Decision Making, 2021, 21(1).

[20] Àvila Callau Aitor, PérezAlbert Yolanda, Serrano Giné David. Quality of GNSS Traces from VGI: A Data Cleaning Method Based on Activity Type and User Experience [J]. ISPRS International Journal of Geo-Information, 2020, 9(12).

[21] Feiyu Lian, Maixia Fu, Xingang Ju. An Improvement of Data Cleaning Method for Grain Big Data Processing Using Task Merging [J]. Journal of Computer and Communications, 2020, 08(03).

[22] Information Technology - Data Systems; Studies from Harbin Institute of Technology Reveal New Findings on Data Systems (A Review of Data Cleaning Methods for Web Information System) [J]. Information Technology, 2020.

[23] Domenico Vitale, Gerardo Fratini, Massimo Bilancia, Giacomo Nicolini, Simone Sabbatini, Dario Papale. A robust data cleaning procedure for eddy covariance flux measurements [J]. Bio geosciences, 2020, 17(6).

[24] Jinlin Wang, Xing Wang, Yuchen Yang, Hongli Zhang, Binxing Fang. A Review of Data Cleaning Methods for Web Information System [J]. CMC: Computers, Materials & Continua, 2020, 62(3).

[25] Science - Analytical Science; Researchers from Southeast University Discuss Findings in Analytical Science (Robust data cleaning methodology using online support vector regression for ultra-short baseline positioning system) [J]. Journal of Engineering, 2020.

A Named Entity Recognition Model Based on Multi-Task Learning and Cascading Pointer Network

Chaoyang Geng
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 541211200@qq.com

Yi Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 168565012@qq.com

Peng Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 643973010@qq.com

Jiejie Zhao
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1838481537@qq.com

***Abstract*—Because, the existing named entity recognition models lack the Specificity of the field, and most of their combine the prediction of entity Location and entity category, which results in the accumulation of errors. So, named entity recognition model based on multi-task learning and pointer network is proposed, and innovations are made in the task Construction and domain entity information utilization in the model of Named entity recognition. This model is based on Transformer with multi-Head attention mechanism, and decomposes traditional tasks in entity recognition tasks and entity classification tasks, and carries out multi-tasks Task learning to reduce the accumulation of errors between tasks. Model in this paper also uses the similarity calculation based on the Comprehensive description of the entity category in entity classification Task for better pertinence of the domain entity. Experiments are conducted on public datasets and domain datasets to prove the advancement of the Model.**

Keywords-Named Entity Recognition; Multi-Tasks Task Learning; Transformer

I. INTRODUCTION

Named Entity Recognition refers to the process of extracting specific words from natural language corpus. Nominated entity recognition tasks also have some problems in entity labeling. Traditional model predictions often use BIO [1] or BIOES labeling methods, which define both entity location and entity category. This means that the prediction of each word by traditional named entity recognition models requires that the categories of named entities and entities be combined. This prediction method has a large problem. If any of the named entity subtags or entity category subtags within an entity predicts errors, the entire entity predicts errors, which can easily lead to the accumulation of errors. Currently, researchers have separated traditional

entity recognition tasks into named entity prediction tasks and entity classification modules. Zheng [2] et al. proposed a named entity recognition algorithm based on Multitask learning. Multitask learning is an integrated learning method [3] [4], which improves multiple tasks by training several tasks at the same time. Based on BiLSTM model [5] [6], the traditional named entity recognition model is divided into two modules: named entity prediction task and entity category classification task. It also uses multitask learning to train, which achieves good results, but there are problems that context information other than named entities is not fully used and comprehensive information of entity categories is not introduced. Ding Yi qi [7] et al. proposed a Chinese named entity perception neural network model based on Multitask learning. The traditional named entity recognition task is divided into named entity perception task and entity classification task, and the loss function is optimized to better identify Chinese entities. However, it has the problem of identifying the beginning and end of an entity with two modules, which leads to inconsistency in training and prediction, and lack of organization in the representation of entity categories. This paper proposes a MTL-NER model (a named entity recognition model based on multi-task learning and cascading pointer network) and an entity labeling method based on How Net [8][9][10] semantics. The traditional named entity recognition task is decomposed into global named entity perception and entity classification. The Recognition calculation method based on entity comprehensive description is introduced into How Net knowledge base to classify entities and improve the recognition effect of the overall model. At the same time, entity category description statements are optimized for specific domains to improve the accuracy of domain entity recognition.

II. MTL-NER MODEL

The model structure proposed in this paper is shown in Figure 1 below, the overall model is mainly composed of five layers: data preprocessing layer, shared feature extraction layer, multi task learning layer and output layer. In the data preprocessing stage, a semantic entity annotation method based on how net is proposed, which annotates named entities and categories, and constructs a sample set in sentence units. Combined with how net knowledge base, the comprehensive description of domain entity category is constructed as the input data of the model. In the model construction, the task of entity recognition and entity classification is based on the shared feature extraction layer for text vectorization and feature extraction. The sentence is encoded through the shared feature layer to obtain the feature vector of the sample, and then the domain named entity prediction is carried out.

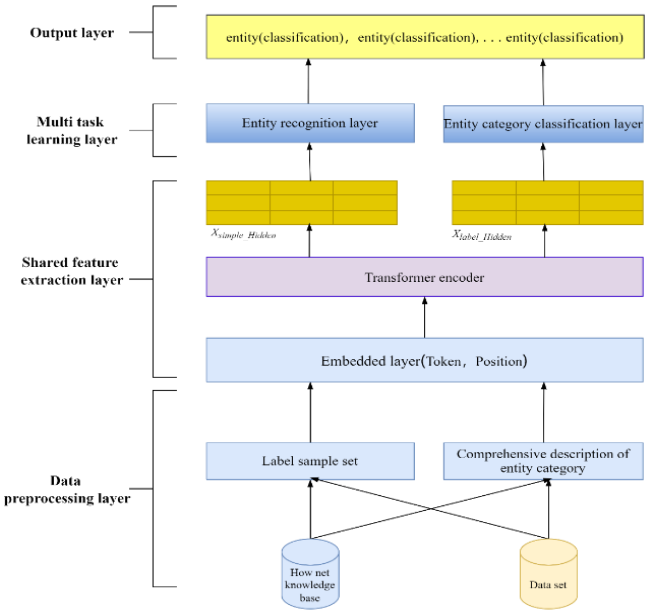


Figure 1. MLT-NER structure diagram

The information of entity classification task comes from the natural language comprehensive description of sample set and entity category, and the same feature extraction layer is used to obtain the feature vector. In the multi task learning layer, the corresponding result vectors are generated from the input feature vectors according to the different tasks. Get the prediction probability results of each task. Finally, the output layer fuses the results of the two tasks, completes the determination of named entities and the classification of entity categories, and obtains the named entity recognition results.

A. Data preprocessing layer

In order to reduce the cumulative impact of traditional entity labeling errors, this paper proposes a new label based on How Net semantics, which labels entity categories and domain entities respectively. In How Net, words are composed of one or more semantics, and each semantics is composed of smaller semantic units (semantics) and dozens of dynamic roles. Figure 2 below is an example:

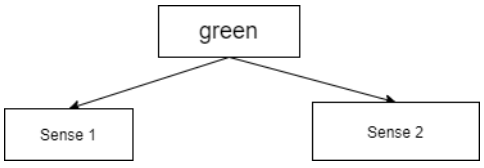


Figure 2. Words and meanings in How Net

The word "green" has two meanings, meaning category 1 is color, and meaning category 2is environmental protection. Then construct labels to classify by meanings. For the input sequence $X = \{x_1, x_1, \ldots x_n\}$, you only need to predict the category, start position, and end position. Based on the above ideas, this topic proposes (Entity category, Start position, End position), taking the domain entity "green" as an example.

Combined with the definition of entity in How Net knowledge base, the meaning item is regarded as the category label of domain entity to improve the accuracy of entity class determination.

Combined with the semantic information of How Net knowledge base, as shown in Figure 4 below, it includes Chinese and English words, semaphores and d DEF_CONCEPT (combination of semantic and dynamic roles) and attributes; Relationships mainly include dynamic roles, hierarchical relationships, and other domain entity category description statements, which are supplemented by Wikipedia and the specific situation of the corpus.

The introduction of the comprehensive description information of the entity category assists the entity classification task, and improves the domain pertinence and classification accuracy of the model. Through the introduction of the entity category description, the ability of the model to obtain the domain information can be improved, so as to strengthen the pertinence of the specific domain. The example of the description statement constructed in this paper is shown in Table 1 below.

TABLE I. EXAMPLE OF COMPREHENSIVE DESCRIPTION OF DOMAIN ENTITY CATEGORIES

Entity Category	**Comprehensive description of entity category**
color	Yellow green blue, environment-friendly color, hue, lightness, saturation and various phenomena of light
Environmental-friendly	Characteristic value, protection, positive evaluation, low carbon, energy conservation and emission reduction, life, agriculture, circular economy, wind and solar power generation
O	General text

B. Shared feature extraction layer

In the shared feature extraction layer, named entity recognition tasks and category classification tasks share an embedding layer and feature extraction layer for joint training. According to the sequence length l, the text $X = \{x_1, x_2, \ldots x_l\}$ is divided and embedded to obtain the input tensor $X_{input} \in \mathbb{R}^{b\times l\times d}$ where b is the batch size, l is the sequence length, and d is the word embedding dimension. Then, according to formula (1) (2), the sequence characters are encoded according to the linear changes of sin and cos functions, and the position vector of the characters in the sentence is obtained $X_{pos} \in \mathbb{R}^{b\times l\times d}$.

$$PE_{(pos,2i)} = \sin\left(pos / 10000^{2i/d_{model}}\right) \quad (1)$$

$$PE_{(pos,2i+1)} = \sin\left(pos / 10000^{2i/d_{model}}\right) \quad (2)$$

Where, pos is the character position, i is the character vector dimension, d_{model} is the hidden dimension of the model. Each encoder is composed of two internal layers: multi head attention mechanism and feedforward neural network. Finally, the input vector of the model is obtained by adding the position code and the word embedded elements:

$$\boldsymbol{X}_{embedding} = \boldsymbol{X}_{input} + \boldsymbol{X}_{pos} \quad (3)$$

In the feature extraction layer, the transformer model based on multi head attention mechanism is used to improve the feature extraction ability of context information. For input vector $X_{embedding}$ decomposes according to formula (4) to obtain query matrix Q, key value matrix K and numerical matrix V. As the input of Transformer encoder module.

$$Q^{(h)},\ K^{(h)},\ V^{(h)} = XW_Q^{(h)},\ XW_K^{(h)},\ XW_V^{(h)} \quad (4)$$

W_Q、W_K and W_V is the weight parameter matrix, $h \in [1, n]$ is the head index, and the head number n is the super parameter. The attention operation is carried out according to formula (5)、(6) obtaining the correlation between each word and other words in the sentence, each word vector contains the vector information of other words related to the current sentence. The result of specific operation is shown in formula (7):

$$Attention(Q,K,V) = Soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

$$head^{(i)} = Attention(Q^{(i)}, K^{(i)}, V^{(i)}) \quad (6)$$

$$MultiHead(Q,K,V) = Concat[head^{(1)};...;head^{(n)}]W^o \quad (7)$$

Then, $MultiHead(Q,K,V)$ and $X_{embedding}$ carries out residual connection to get $X_{attention}$, and normalize the calculation to obtain the standard normal distribution, so as to speed up the training and convergence. The full connection layer feedforward neural network in the encoder is based on $X_{attention}$ is the input, as shown in formula (8) using ReLU as the activation function and performing two linear mappings to complete the expansion and compression of dimensions respectively.

$$FFN(X) = ReLU(XW_1 + b_1)W_2 + b_2 \quad (8)$$

Where, W_1 , W_2 , b_1 and b_2 is the corresponding weight matrix and offset. Finally, $FFN(X)$ and $X_{embedding}$ performs a residual connection and normalization calculation to obtain the output X_{hidden}.

In this paper, the Transformer encoder [11][12][13] based on multi head attention mechanism is used as the feature extraction layer, and the encoder module can be superimposed many times. It realizes unsupervised character level learning and representation of input text sequence under the mechanism of position coding and multi head self-attention.

C. Multi task learning layer

In the named entity recognition task, the shared feature extraction layer extracts the long-distance location dependent features of the context, and outputs the sample vector X_{sample_Hidden}, which contains the feature information of the sample. The structure of the named entity recognition model based on this feature vector is shown in Figure3 below:

The vector representation of the sample word level is obtained by the transformer encoder. In this paper, the cascade pointer network is used to realize the sequence annotation task, that is, two 0/1 sequences are generated by two binary classification networks to determine the start and end boundaries (spans) of entities in the sequence. Each span is determined by a head position pointer (start) and a tail position pointer (end). At the same time, multiple binary classification networks are used for entity recognition.

Each word (token) in the input sequence can be represented as the starting position of an element, and the span composed of any two tokens can be represented as any entity, which solves the problem of nested entity and multi class entity recognition, as shown in Figure 4:

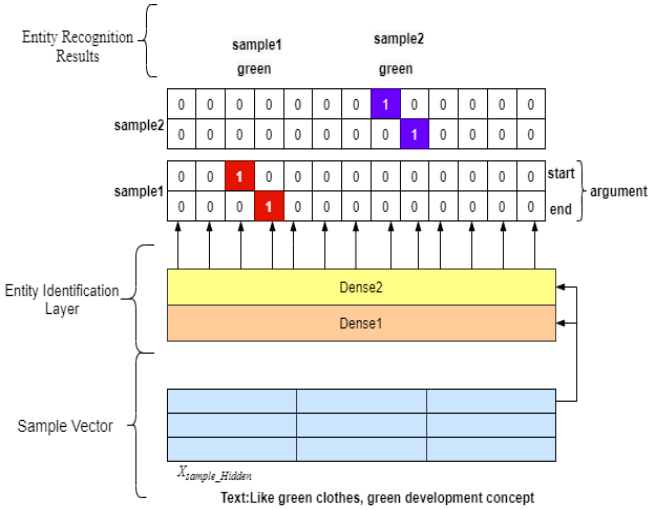


Figure 3. Structure diagram of entity recognition model

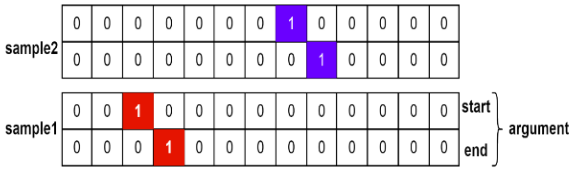


Figure 4. Example of entity recognition

Figure 4 shows the annotation examples corresponding to the input samples. Each entity corresponds to a set of pointer vectors (start, end). By combining the start and end pointer vectors of all entity labels, two-dimensional matrices can be obtained, which are recorded as S_s 、 S_e, That is, each line in S_s and S_e represents an entity type and each column corresponds to a token in the sequence.

In this paper, multiple groups of binary classification networks are used to predict the

possibility that the input sequence is 0/1 at all positions of the start and end pointer vectors corresponding to each entity to determine the start and end positions of the elements. The whole task can be regarded as multi label classification of each token in the input sequence. The probabilities that the i token is predicted as the starting and ending positions of the elements of entity r are $P_i^{s\_r}$ and $P_i^{e\_r}$ respectively, as shown in formulas (9) and (10):

$$P_i^{s\_r} = \sigma\left(W_s^r x_i + b_s^r\right) \tag{9}$$

$$P_i^{e\_r} = \sigma\left(W_e^r x_i + b_e^r\right) \tag{10}$$

Where, $x_i = X_{sample\_Hidden_{[i]}}$, that is, the vector representation of the i token in the input sequence after passing through the encoder, and the superscript s and e are represented as start and end, W_s^r, W_e^r is a trainable weight vector, b_s^r, b_e^r is offset, σ Sigmoid activation letter.

D. Entity classification layer

With the help of How Net knowledge base, this paper improves the construction of domain entity classification and entity comprehensive description, and proposes a similarity calculation model based on entity comprehensive description to output the probability of entity category. The structure of entity classification model is shown in Figure 5.

For the input sample eigenvector X_{sample_Hidden} and entity category comprehensive description information X_{label_Hidden}.This article first introduces X_{label_Hidden} input the full connection layer to realize sentence vector mapping. The formula is as follows:

$$X_{sl} = View\left(X_{label\_Hidden}\right)W_{label} + b_{label} \tag{11}$$

Where, $W_{label} \in \mathbb{R}^{(N+1)\times 1}$, N is the entity category of the sample set. For the sample set with N categories, it is necessary to build N+1 entity category description statements. View refers to the method of reconstructing tensor dimension matrix transformation, transforming the vector with dimension $(N+1)\times d_{label}\times h$ into the dimension of $h\times(N+1)\times d_{label}$, and obtaining the entity category description vector $X_{sl} \in \mathbb{R}^{h\times(N+1)}$. Then the vector dot product is used to calculate the similarity between the sample vector and the N +1 entity category description vector. The calculation formula of entity classification is as follows:

$$C_{x1,x2,..,xd} = sofmax\left(X_{ss} \bullet X_{sl}\right) \tag{12}$$

Finally, the output results of category probability corresponding to each input character are obtained respectively.

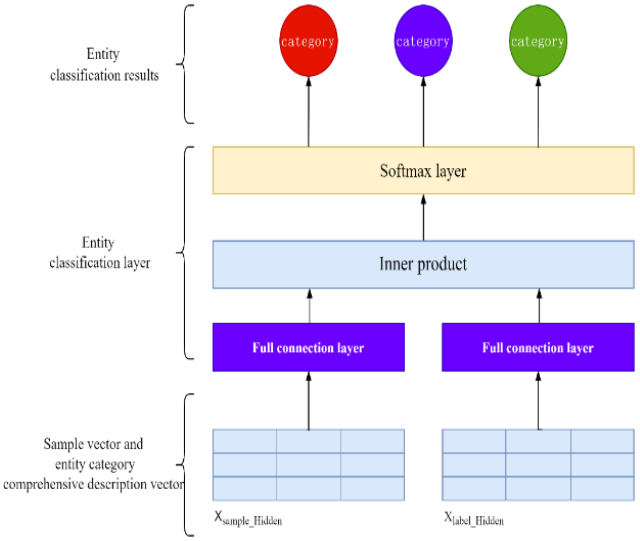


Figure 5. Entity classification model

E. Output layer and loss function calculation

In this paper, the entity recognition probability and entity classification probability are integrated, and the output results are obtained. The start and end positions of entities in the text are marked by the ruler taking method of double pointers, and the entity categories are marked by multiple binary classification networks. Finally, the loss function formula of the entity recognition part is obtained as follows:

$$loss_{sample} = -\sum_{i=o}^{n} y_i^{s-r} \log P_i^{s-r} - \sum_{j=o}^{n} y_j^{s-r} \log P_j^{e-r} \quad (13)$$

Where n is the length of the input sequence; $y_i^{(s_r)}$, y_j^{s-r} is a known correct classification label.

In this paper, the loss function calculation formula based on the entity part of character classification result $C_{X1,X2,\ldots,Xd}$ is as follows:

$$loss_{classfication} = -\sum_{i=1}^{N} C_{x_i} \log(Y_{x_i}) \quad (14)$$

Where, Y_{X_i} entity category label for each character.

Based on the idea of multi task learning, this paper takes the loss function weighting of domain entity recognition task and entity classification as the overall loss function of the model, and the specific formula is as follows:

$$loss = \beta \bullet loss_{sample} + (1-\beta) \bullet loss_{classfication} \quad (15)$$

Among them, $\beta \in [0,1]$ is the super parameter of the model, which converges to the optimal value with training to improve the effect of the model. In order to verify the versatility of the model on different data sets, this paper selects four Chinese named entity public data sets as experimental objects. The data sets include MSRA, ontonotes4.0, cluner2020 and CMeEE, and are compared on four Chinese public data sets to verify the progressiveness of the model in this paper.

III. EXPERIMENT AND A NALYSIS

A. Data set introduction

MSRA is a Chinese dataset released by Microsoft Research Asia, which comes from the news field and is also the benchmark dataset for Chinese named entity recognition. It contains about 90000 Chinese named entities and annotation data. Entity categories include three categories: location, organization and personnel.

Ontonotes4.0 is a Chinese dataset covering multiple data sources. Sources are not limited to telephone conversations, news agencies, radio news, radio conversations and blogs. This article selects four categories of entities in the version, such as people and organizations.

CLUENER2020 data set is a Chinese fine-grained named entity recognition data set, which is based on the open-source text classification data set THUCNEWS, and selects some data for fine-grained annotation. The data set is divided into 10 different categories and 12000 sentences.

The CMeEE dataset originated from chip 2020 (China health information processing Conference). By extracting entities from sentences, they are classified into nine categories: diseases, clinical manifestations, drugs, medical equipment, medical procedures, body, physical examination, microorganisms and departments, with 25000 sentences.

B. Experimental indicators and parameter settings

This paper selects the criteria commonly used in named entity recognition tasks: precision (P), recall (R) and F1 score to evaluate the performance of the model, and selects the result of F1 as the main criterion. The specific calculation formulas of accuracy rate, recall rate and F1 score are as follows:

$$precision = \frac{TP}{FP + TP} \tag{16}$$

$$Recall = \frac{TP}{FN + TP} \tag{17}$$

$$F1 = \frac{2 \times precision \times Recall}{precision + Recall} \tag{18}$$

During the experiment, the parameters are shown in Table 2 below:

TABLE II. EXPERIMENTAL PARAMETERS

Parameter	Value
Optimizer	SGD
Learning rate	5e-6
Activate function	ReLU
Entity category length limit	16
Enter length limit	128
Batch size	6
Deep learning framework	Pytorch
Number of GPUs	1

C. Experimental results and analysis

In this paper, the proposed named entity recognition model MTL-NER based on multi task learning and cascading pointer network is compared with the leading-edge model on the test set of each data set. In the experiment, this paper uses a new ternary tag [entity category, entity start position, end position], while BiLSTM-CRF and BERT-BiLSTM-CRF are trained with traditional BIOES tags, and BERT-MRC uses [entity start position, End position] as a label. In this paper, BiLSTM-CRF is used as the baseline of the experiment. BERT-BiLSTM-CRF is a Transformer based pre training model, which regards the named entity recognition task as a sequence marking task. BERT-MRC [14][15] is a retraining Bert model based on machine reading comprehension (MRC). MTL-NER, a named entity recognition model based on multi task learning and cascading pointer network, achieves the best accuracy, recall and F1 on four Chinese named entity datasets. The F1 of this model is improved by 0.77%, 2.62%, 2.27% and 3.32% respectively compared with the model with the best experimental results on the four Chinese entity public data sets of MSRA, OntoNotes4.0, CLUENER2020 and CMeEE. Compared with the baseline model based on BiLSTM-CRF, it is improved by 16.09%, 35.63%, 22.85% and 24.93% respectively, which is enough to show the

progressiveness of this model. The specific performance of each model is shown in the following table 3:

TABLE III. MSRA DATASET MODEL INDICATORS

MSRA			
Model	P（%）	R（%）	F1（%）
BiLSTM-CRF	87.47	85.23	83.34
BERT-BiLSTM-CRF	95.15	94.85	95.00
BERT-MRC	96.28	95.74	96.01
MTL-NER	97.07	96.43	96.75

TABLE IV. ONTONOTES4.0 DATASET MODEL INDICATORS

OntoNotes4.0			
Model	P（%）	R（%）	F1（%）
BiLSTM-CRF	73.45	60.07	61.71
BERT-BiLSTM-CRF	79.23	79.58	79.40
BERT-MRC	82.49	81.23	81.56
MTL-NER	84.87	82.56	83.70

TABLE V. CLUENER2020 DATASET MODEL INDICATORS

CLUENER2020			
Model	P（%）	R（%）	F1（%）
BiLSTM-CRF	67.23	65.42	66.31
BERT-BiLSTM-CRF	77.42	78.15	77.78
BERT-MRC	79.04	80.26	79.65
MTL-NER	82.14	80.79	81.46

TABLE VI. CMEEE DATASET MODEL INDICATORS

CMeEE			
Model	P（%）	R（%）	F1（%）
BiLSTM-CRF	56.41	49.52	52.74
BERT-BiLSTM-CRF	68.98	66.25	67.59
BERT-MRC	71.26	69.34	70.29
MTL-NER	73.13	70.68	71.89

Because more features are integrated into the word vector generation stage, it can quickly achieve better performance in the training process. However, in order to obtain better results on the test set, it is very important to select the appropriate Dropout, which can prevent the model from over fitting and is more robust. In order to determine the appropriate value of Dropout, this paper conducts several groups of comparative experiments on four data sets in the process of selecting Dropout for named entity recognition, F1 value is mainly used as the measurement standard in the experiment, and the results are shown in the following figure6, figure7, figure8 and figure9:

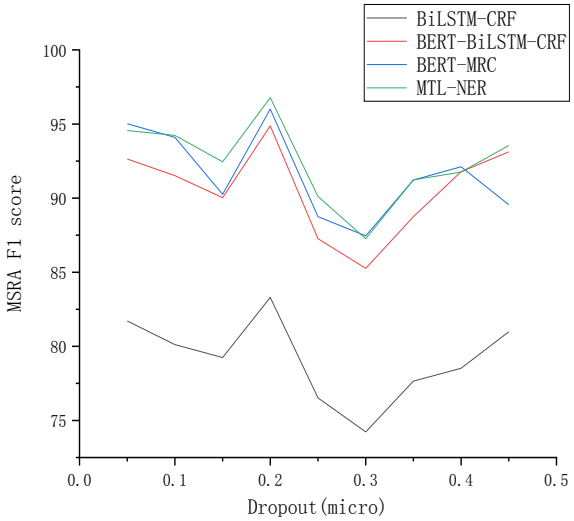


Figure 6. MSRA dataset dropout

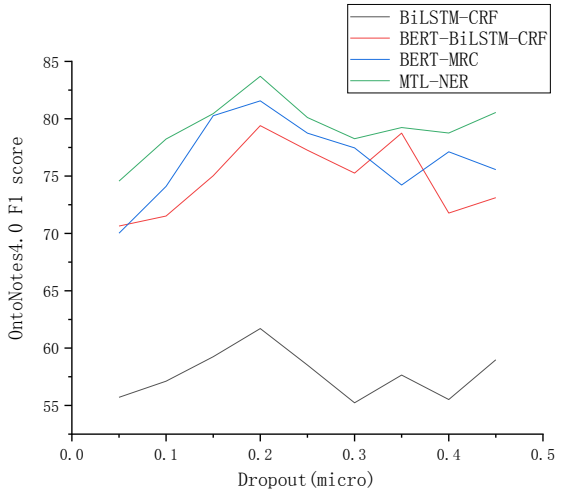


Figure 7. OntoNotes4.0 dataset dropout

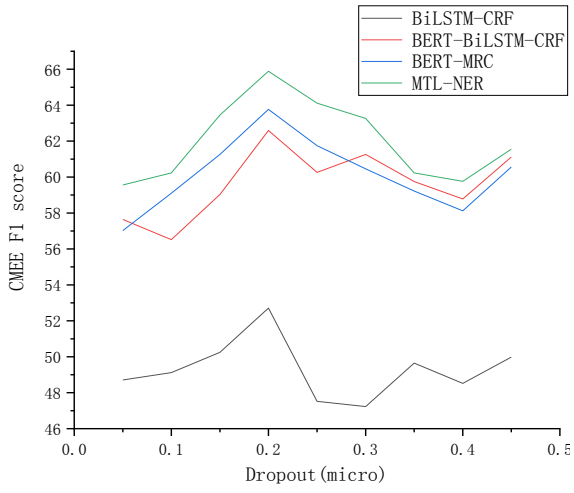


Figure 8. CLUENER2020 dataset dropout

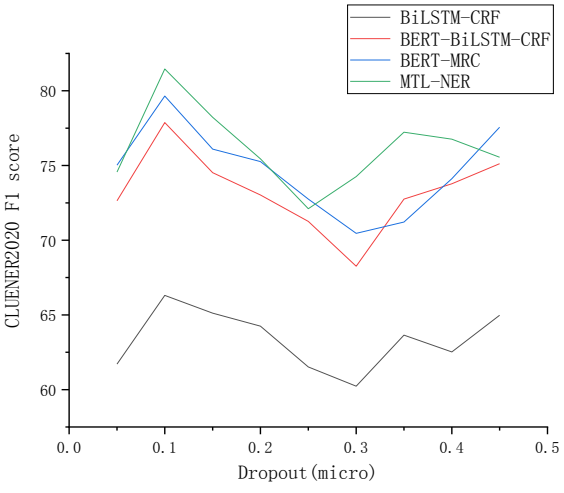


Figure 9. CMeEE dataset dropout

After comparison, in the CLUENER2020 data set, dropout=0.1 is selected, and in MSRA, Onto-Notes4.0 and CMeEE data sets, if dropout=0.2 is selected, the model can achieve the most ideal experimental results. We take the data set of CLUENER2020 data set as an example, and shows the recognition results of domain entities of each model we intercepted:

It is obvious that BiLSTM-CRF model and BERT-BiLSTM-CRF model have poor recognition effect on long text, while BERT-MRC model has multiple recognition errors, and the MTL-NER model proposed in this paper achieves ideal effect on long-span entity recognition.

IV. CONCLUDING REMARKS

In the experiment of named entity recognition model, compared with the advanced model, the MTL-NER model in this paper performs well on four Chinese public data sets, and achieves 96.75% F1 on MSRA data set, which is 0.77% higher than the existing advanced model BERT-MRC. This model performs well on multiple data sets covering multiple fields, which proves the versatility of this model for different fields. The experiment also finds that the model based on Transformer is better than the model based on BiLSTM, which verifies the information extraction and utilization ability of Transformer and the rationality of using Transformer as feature extractor in this paper. Finally, experiment Prove the progressiveness of the innovation model in this paper.

REFERENCE

[1] RAMSHAW L A, MARCUS M P. Text Chunking Using Transformation-Based Learning [J]. Springer Netherlands, 1999, 15-176

[2] Zheng C, Cai Y,, Xu J, et al. A boundary-aware neural model for nested named entity recognition[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).2019:357-366

[3] RUDER S. An overview of multi-task learning in deep neural networks [J]. arXiv per-print arXiv:1706.05098, 2017.

[4] Liu X, He P, Chen W, et al. Multi-Task Deep Neural Networks for Natural Language Understanding[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4487-4496.

[5] CHIU J, NICHOLS E. Named Entity Recognition with Bidirectional LSTM-CNNs[EB/OL]. https://arxiv.org/pdf/1511.08308.pdf, 2021-02-22.

[6] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Australia, 2018:1554-1564.

[7] Ding Yi qi, Yan Dan feng. A neural network model of Chinese named entity boundary perception based on multi task learning [EB / OL], Beijing: China Science and technology papers online [2021-02-22].

[8] Dong Zhen-dong, Dong Qiang, Ebrary Inc. How Net and the computation of meaning［M］. World Scientific, 2006.

[9] Dai Liu-ling, Liu Bin, Xia Yu Ning, et al. Measuring semantic similarity between words using How Net［C］/ /International Conference on Computer Science & Information Technology, IEEE Computer Society, 2008.

[10] Ma Yong-qi, Han De-pei, Meng Li-rong, et al. Lexical semantic similarity algorithm based on How-net [J]. Computer Engineering, 2018, 44(6):151-155.

[11] LI B, KANG X D, ZHANG H L, et al. Named entity recognition in Chinese electronic medical records using transformer-CRF[J].Computer Engineering and Applications, 2020, 56(5):153-159.

[12] YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding [J]. Computer Engineering, 2020, 46(4):40-45.

[13] DONG Z, SHAO R Q, CHEN Y L, et al. Named Entity Recognition in Food Field Based on BERT and Adversarial Training [J]. Computer Science, 2021, 48(5):247-253.

[14] Li X, Feng J, Meng Y, et al. A unified MRC framework for named entity recognition [J]. arXiv preprint arXiv:1910.11476, 2019.

[15] Li X, Yin F, Sun Z, et al. Entity-relation extraction as multi-turn question answering [J]. arXiv preprint arXiv:1905.05529, 2019.

A Datum Schemes Based on TOC Monitoring System

Yuxi Jia
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 562718076@qq.com

Mei Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 7586489@qq.com

***Abstract*—Using the giant parallelism of photons to develop a new type of computer with more processor bits and data, the ternary optical computer, with the birth of the ternary optical computer SD 16 physical machine, the team built a monitoring system structure for task management, It is mainly used to deal with various complex data problems and provide parallel services for computing requests of multiple users. The storage of data in the monitoring system has also become a crucial issue. Based on the existing achievements of the team, this paper designs a data storage scheme for the management software of the ternary optical computer monitoring system, and finally summarizes and looks forward to the next research direction of the ternary optical computer data storage.**

Keywords-Ternary Optical Computer; Task Management; Data Storage

I. INTRODUCTION

With the development of electronic computers and the advent of the 5G network era, in order to break through the limitations of too low information transmission speed and too few data bits for electronic computers, scientists in the field of computing have gradually tried to develop a faster, more reliable Three-valued computer is one of the new types of computers that require high equipment and less power consumption. Its principle is to use the two mutually perpendicular polarization directions of the optical state and the non-optical state to represent information. At present, the relatively complete architecture and physical machines that can be run are the ternary optical computers and ternary electronic computers developed by the ternary computer team of Shanghai University. The development of ternary optical computer has been more than twenty years so far, and a series of considerable results have been achieved. The ternary optical computer is a photoelectric hybrid computer. It is a computer with electrical control and optical operation". It has many operator bits, the operator bits can be arbitrarily grouped, and each group of operator bits can be assigned to different users. The computing function of each bit of the operator can be reconfigured at runtime [1-3]. The ternary electronic computer is researched on the basis of many achievements of the ternary optical computer. The main idea is to transplant the structure of the ternary optical processor SD16 to the processor constructed by the electronic circuit, and realize the multi-value electronic processor. Arbitrary grouping of processor bits can be realized by using the feature of no correlation between the operator and the number of bits of the operator, and the grouping of processor bits can also be used independently.

DOI: 10.2478/ijanmc-2022-0028

In 2010, the ternary optical computer team gradually completed the data splicing technology and data clipping technology [5], monitoring program [7], MSD binary adder[6].These achievements have brought new progress in numerical aspects of ternary optical computers. Among them, the monitoring program enriches the computer monitoring system's management theory and technology for data bits, and paves the way for users to use ternary optical computers conveniently and efficiently [7].

In 2017, SD16, the prototype system of ternary optical computer, completed the 48-bit TW-MSD parallel addition experiment. This achievement marks that the hardware structure of ternary optical computer has been perfected, and the application of ternary optical computer has been explored since then.

One of the characteristics of the ternary optical computer is that the more data it processes, the higher the computing efficiency. The SD16 monitoring system effectively manages the optical computer processor resources with huge data bits. There is no doubt that how to store the data in the ternary computer monitoring system is also the research content of the ternary optical computer at this stage. For a ternary optical computer, the ideal memory is the optical rotatory properties of some new materials to store the polarization direction of light, but no available raw materials have been found so far; another storage method is to store the light itself, "fiber ring memory" is A usable device, with stable storage effects over 30 min has been observed, for the registers of ternary optical computers. Before the appearance of the ideal optical polarization state memory, the existing binary semiconductor memory can be used to store ternary data through group code conversion to form the memory of the ternary optical computer [8]. Therefore, in the operation process of the ternary optical computer, we need to manage the data and solve the following problems: first, in what form is the data stored in the memory; second, what functions should the data storage module have for each SD16 monitoring program? module usage.

II. Related work

A. Computational data model

Figure 1 shows the calculation data model of TOC. The working process of this model is as follows: the user program of the client submits the calculation task to the SZG file receiver/sender (communication server) in the format of SZG file, and TOC task management. The software receives the SZG file, and analyzes the file to obtain the reconstruction information and calculation information of the lower computer, thereby forming a reconstruction frame and a calculation data frame. The TOC large task management software transmits the reconstructed frame information to the lower computer control software, and the lower computer control software parses the reconstructed frame information to form the reconstructed latent image table of the optical processor, which is sent to the reconstruct or to complete the reconstruction. The TOC task management software takes out the calculation data in sequence from the SZG file, and organizes these calculation data into the input data of the composite optical calculator. These sorted data are sent to the composite optical calculator for calculation until all original data calculations are completed, the result beam is sent to the decoder for decoding to obtain binary result data. These result data are organized into calculation result files by the task management software, and the result files are transmitted to the user program of the client through the SZG file receiver. Since TOC will process the data files of different users for calculation, one user corresponds to one SZG

file, TOP will reconstruct the calculator and calculate the data file according to the SZG file, so a space should be opened up to store all the data required by TOP, which should be Contains two worksheets: compound operator data table and operation result data table. The two tables corresponding to each user are different. After TOC processes the operation data of a user, the space occupied by the worksheet should be released in time to increase the memory. space utilization.

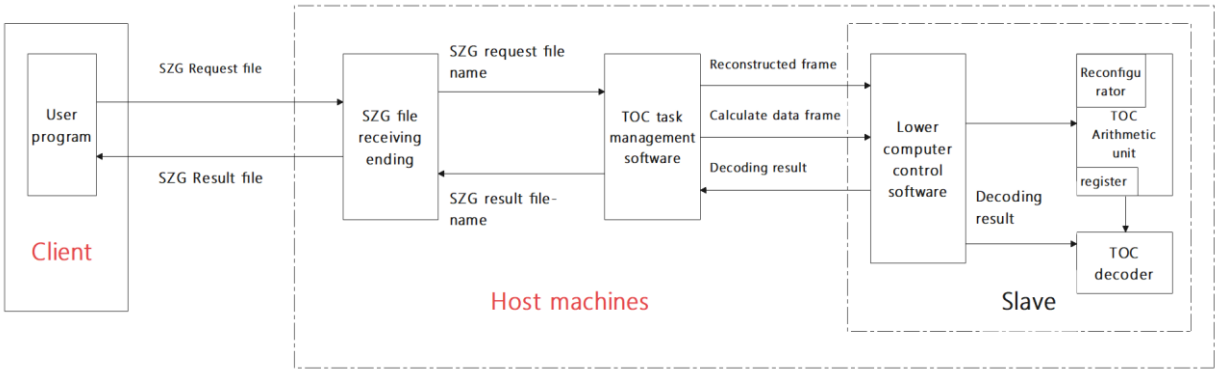


Figure 1. Computation-Data Mode

B. The architecture of ternary optical computer monitoring system

Figure 2 shows the architecture of the ternary optical computer monitoring system, which is mainly divided into the client and the server to complete the work jointly. The client mainly includes user interface, inspection, communication transformation, network communication and communication inverse transformation. The server side mainly includes network communication, data sorting; request scheduling, timer, processor allocation, operator reconstruction, encoder control signal generation, reconstruction register, hardware reconstruction, encoder A, encoder B, Modules such as camera A, camera B, operator, decoding A, decoding B, generating operation results, and generating return results. Some of the modules on the server side are placed on the upper computer and some on the lower computer for the following reasons:

- In order to make the optical processor of the ternary optical computer focus on optical computation. The TOC optical processor is suitable for processing big data and massive data, it can handle multiple tasks at the same time, but there are certain limitations. Assuming that the maximum number of tasks it can process at the same time is TaskMax, when the number of operation requests is greater than TaskMax, there will be operation requests waiting on the host computer. In order to improve the efficiency of the whole system, let the lower computer, namely the optical computer, focus on optical computing, and put the modules of network communication, data sorting, request scheduling, processor allocation, timer and generating return results in the upper computer;

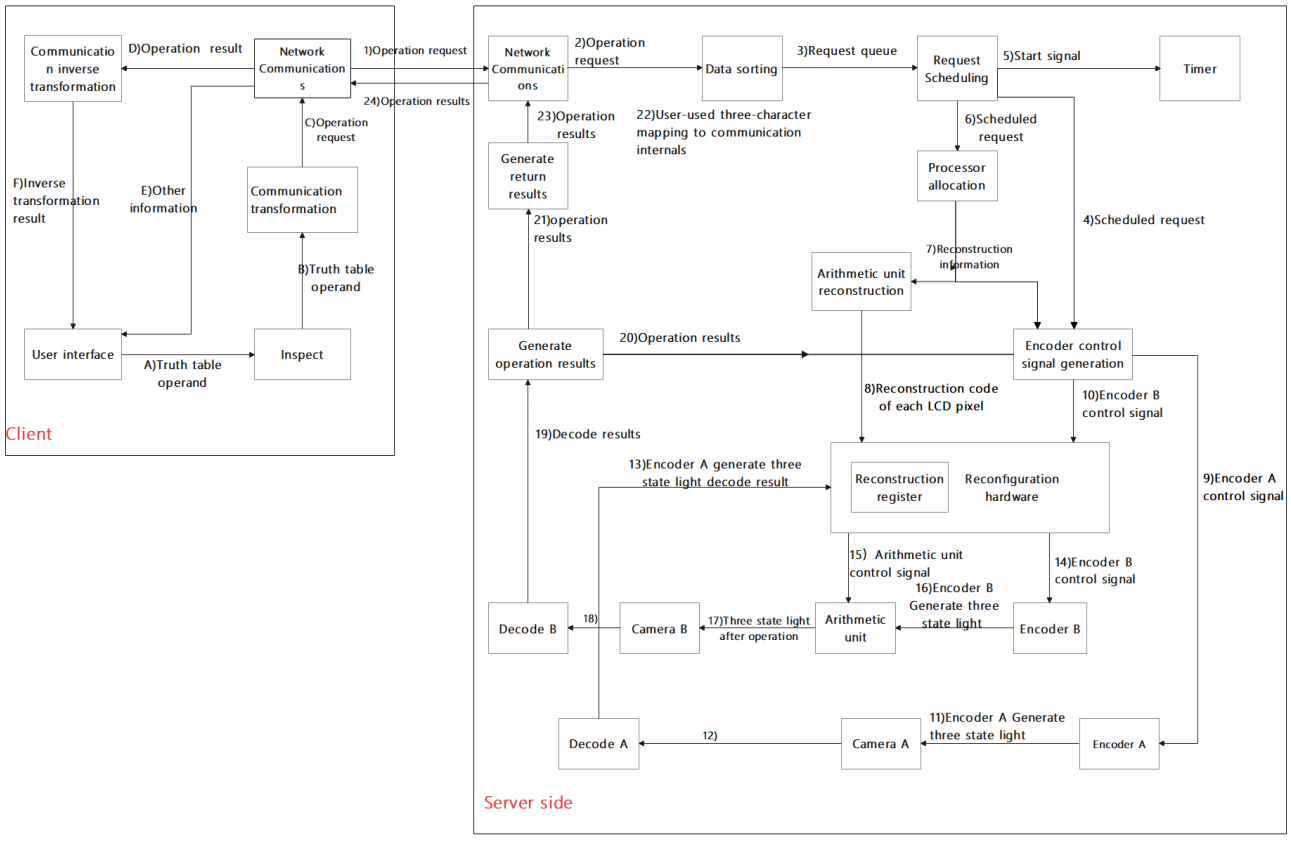


Figure 2. Ternary Optical Computer Monitoring System Architecture

- In order to reduce the amount of communication between the upper computer and the lower computer, the modules for reconstructing the operator, generating the encoder control signal, decoding A, decoding B and generating the operation result are placed in the lower computer. Other optical hardware modules run on the server-side lower computer. In addition, the request scheduling module and the data sorting module can be exchanged with each other, but the efficiency of the system will be different. If the operation request scheduling is performed first, and then the data sorting is performed on the scheduled operation request, the data sorting at this time will occupy a certain time in the "operating" state of the operation request. If according to the figure, the network communication module will organize the data after receiving the user's operation request. For a certain operation request, although the time for data sorting and scheduling is unchanged in these two cases, the data sorting is completed while it is waiting for scheduling and other operations performed by the optical processor, so that the to a certain extent, it provides the overall efficiency of the system.

C. SD16 monitor software

The SD16 monitoring program software comprehensively and clearly describes the overall view of the bottom control software of the ternary optical (electronic) calculator, the functions of the main software modules, the functions of the main

functions, the calling relationship between functions, the sharing and transfer of parameters, The integration method of third-party routines, the cutting method of this software for different purposes, etc. are the key technologies. At the same time, it provides software boundary and framework for the compilation of each function. The system used to study the underlying software and debug application routines of the ternary optical computer should be installed: three modules: administrator operation, online user operation and offline user operation. For the whole program software, if you want to fully experience the system of the ternary optical computer, you only need to install the system that does not include the offline module; and, on the premise that the system is completely stable, the administrator operation module may not be installed.

1) Main program module

The main program is the entrance of the SD16 monitoring program. The user enters the monitoring program, and the main program displays the interface, prompting the user to select the operation to be performed options.

2) Admin action

As mentioned in the introduction of Chapter 1, a ternary optical computer is a electrically controlled, optically operated" computer. In the ternary optical computer monitoring system, the server receives the operation data sent by the client and generates the operation request chain list through the data sorting module, and uses the encoder control signal module to encode the scheduled operation request, and the encoder A module is responsible for transmitting the operation request. The generated tri-state light is given to camera A. The reconstruction hardware module includes reconstruction registers, high-resistance gates, resistors, etc. It needs to receive the code A to generate the three-state light decoding result and the reconstruction code of each liquid crystal pixel to complete the operator reconstruction. In many cases, we need the program to obtain administrator operation authority to perform some operations that require higher authority. For ternary optical computers, higher authority operations include coordinate mapping (scaling) between operator pixels and camera pixels, computing the output light state judgment rule setting (threshold value) of the device can be completed by the administrator.

3) Exit the administrator operation and return to the main program page

This module returns to the previous system page for selection.

4) Calculator pixel and camera pixel scaling

This module is mainly responsible for completing the acquisition of data operation results.

The upper computer in the server side receives and organizes the request data sent by the user through the client into an operation task list, and the lower computer selects an operation task according to the operation request priority of the operation task list, and assigns an operator to complete the data operation. Among them, the data calculated by the operator is not the original data input by the user, but the optical signal encoded by the reconstruction hardware of the lower computer after generating the encoded signal, and the optical signal is transmitted to the operator. After the operator completes the optical data operation, the result is Output to the LCD panel, use the camera to obtain the operation data on the LCD panel, decode the optical data and generate the operation result in the form of ASCII code and return it to the client user interface.

5) Calculator pixel output light state judgment rule setting

This block is used to assign operators to computing tasks.

The characteristics of the ternary optical computer include that the arithmetic unit can be reconfigured on demand to handle massive data. After the lower computer selects the operation request, the reconstruction hardware completes the reconstruction of the arithmetic unit. According to the SZG file, the operation data types received by the lower machine include simple data types and structured data types. When performing data operations, computing tasks with the same computing rules can be performed at the same time, reducing the time for hardware reconstruction. The more data that one operator processes, the higher the computing efficiency. In addition, to test the result field length of each operator, you can choose to note one by one, or you can choose to test continuously, output it to the LCD panel, and obtain the operator field length through the camera. This operation can be used for computing requests.

D. Connected User Operation

For a system for users to experience the characteristics of a ternary optical computer, a module that does not include offline user operations can be installed, and if the machine structure is stable enough, the administrator operation module can also not be installed. The on-line operation of the ternary optical computer refers to connecting the client, the user inputs the operation data to the lower computer on the server side through the client platform, and the lower computer then performs complex data operations. The user enters the connected operation interface through the main program, and the host computer prompts the user to select the operation to be performed. The specific operations include:

1) Quit

Returning to the main program module, the user re-enters the operation to be selected.

2) The user selects an existing compound operator to perform data operations

One of the characteristics of the ternary optical computer is that the server needs to reconstruct the calculator at any time according to the data calculation problem sent by the client. Therefore, the SD16 monitoring program software needs to develop a worksheet to save the defined compound calculator. A specific function is required to append records to the data in the data table, allowing changes and timely updating of the data in the worksheet. The user operates the existing compound operator according to the data table of the compound operator displayed by the host computer. If the structure of the compound operator in the data table needs to be changed, the server-side program saves the changed operator structure; otherwise, the compound operator is used. Complete the data operation, and save the structure of the called compound operator into the corresponding structure array. After the user completes the selection, return to the previous operation.

3) The user chooses to create a new compound operator to perform data operations

The computer-connected user operation is a module that helps users experience the characteristics of a ternary optical computer. After inputting the operation data, the user can choose to create a new compound operator to complete the operation of the data. The user inputs the operator number of the user-defined composite operator, and the reconstruction hardware module performs operator reconstruction after receiving the user request to complete the reconstruction of the operator. At this time, the operator number in the composite operator data worksheet is updated. After the update is complete, select Continue

Refactoring Component or End Refactoring Component.

4) Enter raw data

This module also contains the following two operations:

a) Replacing the compound operator

As mentioned above, the on-machine operation helps the user experience the characteristics of the ternary optical computer and is user-oriented. Therefore, the user can also choose to replace the composite operator to complete the data operation.

b) Input data

After the user completes the selection of the operator, the data is input, and the input methods available for the user to choose include single-group input data and multi-value input data. The input of a single set of data means that the user inputs the calculated data multiple times without repetition into the structured data array, which is used to store the single structured data that the user needs to calculate; the other is multi-valued data input, the user needs to input continuously. The data value of the operation, the data value corresponds to the compound operator name one-to-one, and the data value is stored in the structure array. After the data input is completed, update the value of the composite operator.

c) Rebuild Optical Processor

In the ternary optical computer monitoring system, to complete the user operation, it is necessary to allocate processor resources for each operation request, and the processor resources include an encoder, a decoder and an operator. The processor allocation module on the server side allocates optical processor resources to the user according to the calculation amount and priority of the user's calculation request. A good processor allocation algorithm can improve the overall performance of the system. For a ternary optical computer, the processor resource is the allocation of liquid crystal pixels, that is, the liquid crystal pixels must be addressed before the operation. After the processor bit allocation is completed, each processor bit corresponds to the reconstruction instruction of the relevant operation rule, the task management software first generates the reconstruction latent image and then completes the reconstruction of the processor. In the SD16 monitoring program system operation, the system searches for the name of the compound operator given by the user. If it is found, it calls the function to generate the reconstructed latent image array to obtain the reconstructed data, and then calls the reconstruction function to complete the reconstruction operation of the optical processing and returns to the above primary operation. If the system cannot search for the compound operator name entered by the user, the user will re-enter it.

5) Get the calculation result (Camera B)

After the user inputs the operation data, the server performs the operation of the data, and finally returns the operation result to the client. There are two actions for the user to choose from in this module:

a) Confirm compound operator

Wait for the user to confirm the compound operator and then perform data operation.

b) Replacing the compound operator

Displays the composite operator structure table, and the user inputs the value number of the composite operator to be selected.

The most significant difference between an optical computer and an electronic computer is that an optical computer uses the characteristics of light to complete data operations, and the final calculation result is also optical data. Therefore, the operator maps the resulting optical data to the

liquid crystal panel after completing the data operation, and then captures the pixel value of the liquid crystal panel through the camera, decodes the pixel value of the liquid crystal panel into an ASCII code value and returns it to the client.

E. Offline User Actions

Systems used to work with SZG files can only install offline user modules. The offline module can be completely separated from the client side of the SD16 program software. The server side directly receives the data sent by the SZG file generation software platform, performs data operations and operations, and finally returns the operation results to the user.

F. Related data structures

In order to save the user-defined compound operator, original data and corresponding calculation results, the SD16 monitoring program software needs to use the compound operator as the associated keyword, and the function should have the ability to add additional records to the data table. Basic functions such as data exchange, retrieval, deletion, and copying with arrays can improve the data calling efficiency of each module of SD16 monitoring software.

1) Operation request

In the monitoring system, the client sends an operation request containing data such as operands and truth tables to the server, and the operation request has two attributes: priority and operation amount. The operation request does not communicate with other operation requests in the whole calculation operation, so the operation request is independent. According to the input interface displayed by the client to the user, the user should input the following data: "E-mail", "password", "payment level", "domain number Y" and some related information. The format of the operation request data is shown in Figure 3, where the numbers in parentheses represent the number of bits of data. In the following description, the priority of the operation request is jointly determined according to the amount of data requested by the user and the payment fee. The more complex the amount of operation data and the higher the payment fee, the higher the priority number of the operation request.

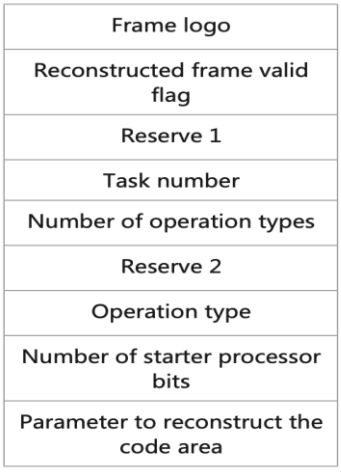


Figure 3. Reconstructed data frame format

2) Reconstruct the dataframe

The operator of the ternary optical computer is composed of several simple primitives. After the processor bit allocation is completed, each processor bit corresponds to the reconstruction instruction of the relevant operation rule. The reconstruction format frame is shown in Figure 4: Frame type flag: It is identified by a data bit. The value of this bit is 1 for reconstructed frame and 0 for data frame.

Reconstructed frame type valid flag: 1 bit, the value of this bit is 1 means the reconstructed frame is valid, 0 means invalid.

Reserved: 6 bits, which are used as extension bits for subsequent development, and are also composed of frame type valid flag and frame type flag

One byte is convenient for task management software and lower computer control software to analyze. Reserved bits are currently set to all 0s.

Internal task number: corresponding to the variable ID, the number of the SZG file in all tasks.

Number of operation types: It means that the SZG file contains several types of operations.

Operation Type: The corresponding calculation type in the SZG file.

Start Bit Order: The start bit on the optical processor of a calculation type in the SZG file.

Number of processor bits: The total number of processor bits required for each operation.

Parameter area: Indicate several parameters of this type of transformation.

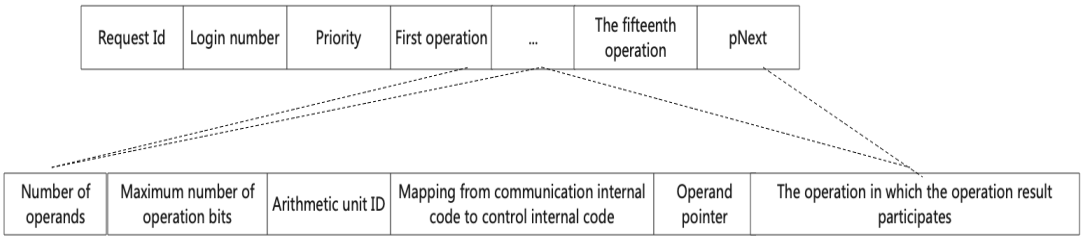


Figure 4. Operation request data format

Reconstruction code area: used to store the reconstruction code of each operation in the SZG file. Frame type flag, reconstructed frame type

The valid flag bit and task number are determined by the task management software according to the actual situation of the task calculation. The number of operation types is obtained by parsing the number of calculation marks in the SZG file, and the number of operation types is obtained by parsing the

3) Calculator generating function

a) Function and calling timing: In the architecture of the ternary optical computer monitoring system, after the lower computer receives the operation request sent by the upper recorder, it now searches the compound operator table for the compound operator that meets the calculation rules of the operation request. Allocate an operator for the calculation request. If it is not found or during the operation of the connected user, the user chooses to create a new operator to experience the characteristics of the ternary optical computer.

b) Step:

Step 1: First determine the operation rules of the operation request (binary logic, three-valued logic, MSD addition, decimal addition);

Step 2: When the operation rule is a binary logic operation or a ternary logic operation, subdivide the operation rule into not, NAND, NOR, OR, AND, XOR operation, and input two or three English words mu as logical symbols;

Step 3: When the user inputs data on the client side, it is in the form of a truth table, and after determining the value in the truth table, determine the maximum number of digits of the data (determine the maximum number of digits that the operator can process);

Step 4: Add the newly created calculator to the composite calculator data table, and increase the number of calculators in the monitoring system by one.

4) Reconstructed latent image group generating

function

This function is used to generate the reconstructed latent image group of the composite operator, and to complete the operator reconstruction, first initialize a structured array to store the reconstructed latent image data. The steps to generate reconstructed latent image data are as follows:

a) Receive operation data: After the server receives the operation data sent by the client, the host computer mounts the operation request on the task list;

b) Select a computing task: the lower computer selects a computing task according to the priority of the computing request, or when the user experiences the ternary optical computer to choose the computing device reconstruction by itself, and analyzes the computing task of the user;

c) Generate reconstruction instructions: generate operator reconstruction instructions through the encoder control signal module to analyze the amount of operation in the operation task and the operation rules;

d) Grouping of latent image values: According to the operation rules (two-valued logic, three-valued logic, etc.), the reconstruction instructions are grouped and stored, and the reconstructed values are mapped to the LCD panel;

e) Hardware reconstruction: The reconstruction register uses the reconstruction code of the liquid crystal pixel to complete the reconstruction of the arithmetic unit.

III. Problem Description

Any computing task running on an optical computer or an electronic computer is inseparable from the operating system's task scheduling and memory allocation. The ternary optical computer needs to deal with huge amounts of data. Effective management of the operation data is one of the solutions to improve the computing efficiency of the ternary optical computer.

Virtual memory is proposed to solve the problem that the existing memory cannot meet the memory needs of a large process, and it is one of the technologies often used by computer operating systems. Each process has an independent logical address space, and the memory is divided into several "pages" of size. Each page is a continuous address. For a process, there are many logical memory spaces, some of which correspond to physical A piece of memory, and a part of the unloaded memory corresponds to the hard disk.

The continuous single allocation memory management method is used to manage physical memory, and is mainly divided into single continuous storage management and partitioned storage management.

A. Single Continuous Storage Management

The single continuous storage management method divides the memory into two parts: the system area and the user area; the application program is loaded into the user area, and the application program can use all the space of the user. This method is suitable for a single user and a single operating system. Although it is easy to manage, once all the programs are loaded into the memory, it means that the programs that are not frequently used will also occupy part of the memory.

B. Partitioned storage management

Partition management is proposed to support multiprogramming systems and time-sharing operating systems, and multiple programs are executed concurrently. The commonly used method is called memory compaction. Partitioned storage management divides memory into a

number of equal or unequal sized partitions, one of which is occupied by the operating system and the rest used by applications, each application occupying one or several partitions. Although partitioned storage management can support concurrency, it is difficult to share memory partitions. Partitioned storage management introduces new problems: internal fragmentation and external fragmentation.

1) Fixed partition

The characteristic of fixed partition is that the memory area F is divided into several consecutive partitions of equal size. Although this kind of general is easy to implement, it will cause waste of memory fragmentation and limit the number of programs.

2) Dynamic Partitioning

The characteristic of dynamic partition is to create partition dynamically. When the program is loaded, the operating system divides the program into batches or changes the size of the partition according to the initial requirements. The idea of the algorithm of the too partition is to find a suitable free partition for a program. If the free partition is larger than the original need, it will be divided into two parts, one of which is marked as occupied, and the other is marked as free. The order of partition allocation is usually from the low end of the memory to the high end. One thing to pay special attention to about dynamic rage release is to merge adjacent free regions into larger memory regions. Commonly used partition allocation methods mainly include: first-fit method, first-fit algorithm, best- fit algorithm, and worst-fit algorithm.

IV. ALGORITHM DESIGN

A. Data Classification

In the monitoring architecture, the data is divided into operation data and operation data. The user submits the data operation problem to the TOC in the form of SZG file, because the SZG file data submitted by the user is uncertain. According to the large number of ternary optical processors, in general, the task management software classifies tasks according to the SZG file data format according to a certain strategy, generates different task queues, operates on different task queues, and improves the performance of the ternary optical computer. Overall efficiency, task division strategies are roughly divided into the following categories: calculation type-based strategy, priority-based strategy, payment-based strategy, business behavior-based strategy, and data-volume-based strategy. Strategy: divide computing tasks into large tasks and micro tasks. Operation data is divided into: operation data, operation rules, reconstruction instructions, compound operator table and operation results.

B. Allocate memory space

1) Big task

For the entire workflow of the TOC monitoring system, after the server has sorted out the operation requests sent by the client, the partner algorithm allocates free memory blocks for a task, and creates a dynamic linked list to store some operands, Variables such as operation rules and priorities [12].

For the large task of the ternary optical computer: a large task occupies an optical processor, this strategy can completely avoid reorganizing the original data and extra data, and also avoid the invalid state of the TOP part of the partition, improving the system efficiency ; Construct calculators with the same working time or the same number of iterations as a composite operator to manage the data input and result output of this group of calculators in a unified manner, which can simplify the complexity of

data scheduling; In other words, the more calculators each compound operator contains, the less the number of compound operators, and the easier it is to manage them in groups; in addition, the number of bits of the ternary optical processor can be allocated according to the compound operators , but considering that the number of bits of the composite operator is much smaller than that of the optical processor, multiple composite operators can be constructed on the optical processor, including constructing different composite operators at different positions of the optical processor In order to improve the overall efficiency, it is stipulated that the optical processor should process a large task uninterrupted; according to the ternary optical processor The computer uses an electronic computer as an auxiliary PC. It can be known that the memory of the PC is limited, and all the data in a SZG file cannot be preprocessed at the same time. The data can only be divided into blocks according to the size of the PC memory, and the PC reads a piece of data at a time. Processing, the operand code generated after processing the data is sent to the TOP encoder for data operation; according to the SZG file format, the calculation of a SZG file data can contain different quantities, when a calculator in the compound operator When there are more calculators than other calculators, some calculators have already played with the calculator, while other calculators are still waiting for the data, which will cause the event to idle and reduce the work efficiency. To refine the specific structure of the composite processor, the idling of the entire composite operator is minimized.

There is a large task T. It is known that a large task SZG file can store 1TB of data. The process of allocating a memory block of order to T by the partner algorithm is as follows: In the partner system, all memory blocks are 2 The system searches for a free memory block larger than T in the free_area array. The upper limit of the number of free_area elements is MAX_ORDER, and a variable current is defined to record the order of the currently found free memory block. If the system finds that the subscript of the array element is MAX_ORDER-1 and cannot be found, the erase search fails, indicating that the allocation requirements of task T cannot be met, and a null pointer is returned, indicating that the memory allocation failed; if the search succeeds, the order of the found Assign it to current, remove the home page descriptor of the first free memory block from the current free list, the pointer page points to the array element, reduce the number of the current element free list by one (free_area [current]. nr_free), set one the variable m records this value. If current is greater than order, it means that the memory block is larger than the requested memory block and needs to be divided twice, and the current-order memory block is divided into two current-1-order memory blocks, using B L and B H respectively records the two memory blocks located at the low address and the high address, using the page descriptor pointer to point to B L , and using page+2current to point to the page descriptor of BH; insert BH into the idle current-1 level In the linked list, set the m value of current-1 order plus one, modify the page descriptor of B H by page+2 current+1, calculate the value of current-1, if it is still greater than order, continue to divide until the remaining memory blocks The order is equal to order , point the page pointer to the home page descriptor of the remaining memory block, and subtract 2 from the entire page frame number to complete the memory block allocation of task T.

2) Microtasks

The data calculation amount of microtasks will not exceed that of large tasks, so if there are multiple microtasks with similar calculation rules, they can share the same compound operator or the

calculator in the corresponding operator, reducing the number of reconstructions of the operator. Completing the calculation of multiple tasks within a time, reducing event idling, and improving system work efficiency; if the calculation rules of micro-tasks are different, it is necessary to reconstruct the calculator, but due to the relatively small amount of tasks, the calculation speed will be higher than that of large tasks. But for a ternary optical computer, the larger the amount of data, the more tasks, the higher the work efficiency, and the more the advantages of the ternary optical computer can be demonstrated. Suppose that the SD16 monitoring program receives two micro-tasks T1 and T2 from the client, and now allocates memory blocks of order1 and order2 for them. Obviously, the amount of data in the SZG file of a microtask must be smaller than that of a large task. As mentioned above, the same operator can be used to complete the data operation of multiple microtasks at the same time. Therefore, for the memory allocation of microtasks, improve system resource utilization by allocating a large block of memory for multiple microtasks. The process of allocating space is similar to that of large tasks, except that the size of the free memory block to be searched in the dynamic linked list should be greater than the sum of order1 and order2, and then it is divided. Update the number of free lists in the free_area array in time.

3) Working principle

For the TOC monitoring system, the type of user data, the amount of data, and the calculation rules received by the server may be very different each time. If fixed partition allocation is adopted, the memory space allocated to the task will be insufficient or there is still space left(internal fragmentation),resulting in waste of resources; while dynamic partition allocation may result in external fragmentation and waste of resources.

As mentioned above, fixed partitions limit the number of processes to execute, and when the process size does not match the memory space, the utilization of the memory space will be greatly reduced. The dynamic partitioning method has complex algorithms, high system overhead, and external fragmentation. The partner algorithm can compromise the deficiencies of the above two methods.

Create an array in the buddy system to keep track of free memory blocks. Suppose now that 2 pages of memory are to be allocated, the partner system will look in the book to see if there is free memory larger than required. If there is, it will be divided into two parts, and the allocated memory block will be divided into two parts according to the power of 2. Array free list.

C. Related data structures

The Linux kernel divides physical memory management objects into three levels: nodes, regions, and page frames. Organize these three data structures together; the global array mem_map stores the page descriptor, and the pointer zone_mem_map is created in the zone descriptor to point to the starting page frame page descriptor of the corresponding memory area. Concatenate all node descriptors into a linked list, creating a pointer to the first node descriptor.

Create a free_aera array to track free memory blocks. The array has a maximum of max_order elements. Each element contains the head node of the free list (freee_list) and the length of the free list (nr_free) two attributes; the top of Figure 5 for an array that records the free memory as soon as possible, the following is the memory usage. The data sorting module on the server side of the TOC monitoring system sorts the operation request (Figure 4) sent by the client to generate the request scheduling operation request, and then inserts it into the operation request list according

to the priority. The meeting algorithm allocates space for a request based on its size.

D. Memory block recycling

The buddy system refers to two free blocks of equal size, adjacent, that can be synthesized into higher-order memory blocks as buddies. During the merging process, first determine that the memory blocks reclaimed with equal bands are partners with each other. If the partners are in an idle state, the two are merged into a higher-order memory block, and the process is repeated until the merged memory block has reached the highest order. The merger of partners should also be discussed in categories: partners on the left, partners on the right.

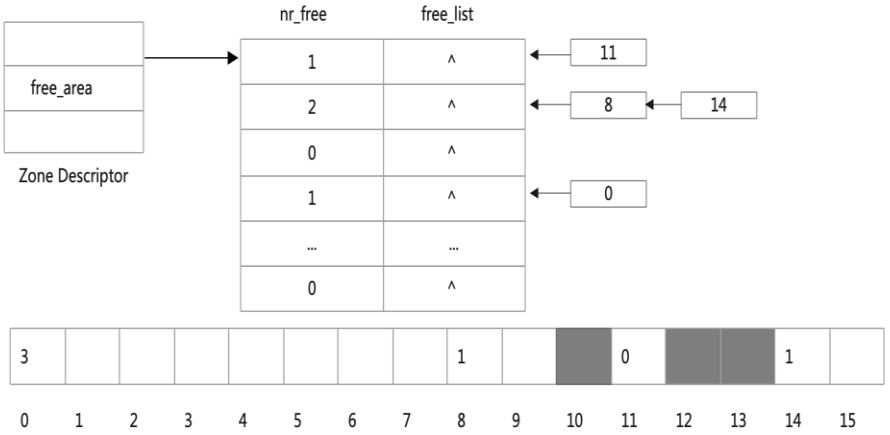


Figure 5. Idle Chain List and Memory Block Distribution Diagram

- The partner is on the left:

$$b = p - 2^k \tag{1}$$

The index of the first page of the k-1 memory block formed by the merger is b and it satisfies that b is divisible by 2k+1, so the lowest k+1 bits of the binary form of b are all 0, so b[k]= 0, set b[k] to 1 to get p, so p[k]=1,

$$p' = p \tag{2}$$

- The partner is on the right:

$$b = p + 2^k \tag{3}$$

The index of the first page of the memory block of order k+1 is merged to form p, and p is divisible by 2k, so the lowest k+1 bits of the binary form of p are all 0, so p[k]=0. Among them, p is an integer, p[k] represents the bit k in the binary form of p,

$$p' = b \tag{4}$$

According to

$$b = p \oplus k \tag{5}$$

The calculation of the partner home page index, the merged high-order memory block home page index can be calculated, and then the partners can be merged.

$$p' = p \wedge b \tag{6}$$

E. Data storage method

The research content is based on the architecture of the ternary optical computer monitoring system, in which the operation data is stored in the form of SZG files, and the operation data is stored in the form of a data table. The developed two worksheets include the original data and the corresponding calculation results, a user-defined compound operator. The two tables are designed with a compound operator named associative key.

- The compound operator table includes the compound operator name, the compound operator structure, the last used time, the calculators contained in the compound operator, the calculator calculation rules, and the compound operator's same name serial number (convenient for users to define the same name operator)
- Original data: In the online user operation, the user can create a compound operator by himself, so it is necessary to define an array to store the operator used for the operation data (the array element is a structured array), and the subscript is the operator bit sequence. Each bit of data input by the user is regarded as a character, and each character is converted into a communication internal code (01, 00, 10) according to the user's operation rules (two-valued logic, three-valued logic, MSD addition);
- Operation result: The composite operator returns the optical data after completing the data operation. After decoding the optical data, it outputs the ASCII code data and returns it to the client user interface. The operation data includes: the number of generated liquid crystal pixels, the operation result pointer, the operation to be involved in the operation result to form a sequence table, the field length (the number of bytes occupied by the operation result), and the operation result data.

F. Data call timing

1) Operation data

After receiving the operation request sent by the client, the upper record of the client sorts it out, and finally mounts it on the operation request linked list, and then completes the task scheduling according to the priority timing scheduling algorithm, selects an operation request and sends it to the next crisis, the lower computer The operation data is analyzed to generate corresponding reconstruction instructions and encoded signals.

2) Algorithm

The lower computer performs hardware reconstruction according to the operation rules of the operation request, and allocates processor resources for the operation tasks. The processor resources include an operator, a reconstruction register, and an encoder.

3) Refactoring instructions

The ternary optical computer will reconstruct the operator according to the user's calculation request. In the online user operation, the user creates a compound operator name to experience the characteristics of the ternary optical computer. After receiving the compound operator name created by the user, the reconstruction register generates reconstruction instructions according to the user's data operation rules, and reconstructs the hardware module. Call the refactoring instruction to complete the refactoring of the operator.

4) Compound Operator Table

According to the operation request sent by the upper computer, the lower computer searches the compound operator table for the compound

operator that meets the operation rules of the operation request; or in the operation of the connected computer user, the user creates the compound operator to complete the calculation of the operation task, and the creation is completed. Then update the data in the compound operator table.

5) Operation result

SD16 monitoring system will first encode the operation task into optical data, and perform operation on the optical data. After the optical data operation is completed, the camera will be used to collect information. The decoding module generates the operation result in the form of ASCII code, and finally inserts the operation result into the operation result linked list, and retrieves the operation result that has completed the operation and returns it to the client for display to the user. If the operation task has not been completed, call the existing operation result to complete the operation when the next operation is performed.

V. SUMMARY AND OUTLOOK

A. Summary

The data management of ternary computer monitoring program software should have the following functions: save the operation data and operation data in the monitoring program, save the user-defined compound operator, original data and corresponding operation results in the form of a data table. The data table file is tracked and updated in real time, and data is exchanged with the array, including functions such as retrieval, deletion, copying, and modification, which are provided to each module call of the monitoring program software.

B. Outlook

As a new type of computer, the ternary optical computer has gradually improved its theoretical system structure and entered the practical stage after nearly 20 years of research and development, and gradually entered people's field of vision. How to effectively manage the computer data that has a large number of data bits in a ternary computer and can be allocated by the processor on demand is also a problem that many scholars are paying attention to at this stage. Effective data storage can improve the computing efficiency of the ternary optical computer. This paper the research content is based on the many research results of the research team. For the SD16 monitoring program software that has been developed, this paper is the first to study the optical computer data management and propose a set of management plans. The data is classified and stored, and then a reasonable storage space is allocated for the data, and the data is updated in time to be called by each module of the monitoring program software, and the space is released in time after the data is used. According to the research tasks at this stage, there are still some imperfections in this paper. In the future use process, the details of the scheme will be gradually improved, and new technologies and strategies will be explored.

REFERENCE

[1] Jin Yi, He Huacan, Lv Yangtian.Basic principle of ternary optical computer [J].Science China E Series:Technical Science, 2003(02):111-115 .

[2] Li Mei. A Review of Ternary Optical Computer Research [J]. Electronic Design Engineering, 2014,22(17):22-25.

[3] Sun Hao, Jin Yi, Yan Junyong. Experimental research on the principle of ternary optical computer encoder and decoder [J]. Computer Engineering and Applications, 2004(16):82-83+136.

[4] Wang Xianchao. Task management and theoretical research of ternary optical computer monitoring system [D]. Shanghai University, 2011.

[5] Li Shuang. Theory and Design of Ternary Computer Programming Platform [D]. Shanghai University, 2019.

[6] Jin Yi, Xu Qun, Ouyang Shan, Han Yuexing, Li Weimin. Structural Quantity Computer - Application Characteristics of Ternary Optical Computer [J]. Science in China: Information Science, 2016, 46(03): 311-324.

[7] Jin Yi, Ouyang Shan, Song Kai, Shen Yunfu, Peng Junjie, Liu Xuemin. Data Bit Management Theory and

Technology of Ternary Optical Processor [J].China Science(Information Science),2013,43(03):361-373.

[8] YunFu Shen,Lei Pan. Principle of a one-step MSD adder for a ternary optical computer[J]. Science China Information Sciences,2014,57(1).

[9] Jin Yi, Gu Yingying, Zuo Kaizhong. Theory, Technology and Implementation of Ternary Optical Computer Decoder [J]. Science in China: Information Science, 2013, 43(02): 275-286.

[10] Jin Yi, He Huacan, Ai Lirong. The principle of carry-to-parallel ternary optical computer adder [J]. Chinese Science Series E: Information Science, 2004(08):930-938.

[11] Huang Weigang, Jin Yi, Ai Lirong, Yan Junyong, Sun Hao. Design and construction of one-hundred-bit encoder for ternary optical computer [J]. Computer Engineering and Science, 2006(04):139-142.

[12] Huang Weigang. Design and implementation of one-hundred-bit encoder for ternary optical computer [D]. Northwestern Polytechnical University, 2005.

[13] Yan Junyong, Jin Yi, Zuo Kaizhong. The Design Theory of Non-advanced (Borrowed) Bit Calculator and Its Application in Ternary Optical Computers [J]. Science in China (Series E: Information Science), 2008 (12):2112-2122.

[14] Zhan Xiaoqi, Peng Junjie, Jin Yi, Wang Xianchao. A Static Allocation Strategy for Data Bit Resources of Ternary Optical Computers [J]. Journal of Shanghai University (Natural Science Edition), 2009,15(05):528-533.

[15] Li Mei, He Huacan, Jin Yi, etc. An optical method for realizing balanced ternary vector-matrix multiplication [J]. Computer Application Research, 2009, 26(10): 3812-3814.

[16] Junjie Peng,Xianshun Ping. An optical implementation method for symmetric MSD number[J]. Optik,2017,143:

[17] Chen Liping, Jiang Jiabao, Liu Yong. A review of research on ternary optical processors [J]. Computer Knowledge and Technology, 2018,14(25):263-264+266.

[18] Li Shuang, Jin Yi, Liu Yuejun, etc. Initial SZG file generation software for ternary optical computer [J]. Journal of Shanghai University (Natural Science Edition), 2018,24(02):181-191.

[19] Li Mei. Design and Implementation of Ternary Optical Computer Experiment System [J]. Computer Technology and Development, 2016,26(10):192-195.

SSD Object Detection Algorithm Based on Feature Fusion and Channel Attention

Leilei Fan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 2547462712@qq.com

Zhiyi Hu
Engineering Design Institute
Army Research Loboratory
Beijing, 100042, China
E-mail: 18992899862@163.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: yujun@xatu.edu.cn

***Abstract*—Aiming at the problems of low object detection accuracy due to complex background and insufficient semantic information of shallow features in the object detection SSD algorithm, this paper improves the existing SSD algorithm. First, the original vgg16 network is replaced by the ResNet50 network, and the residual network structure as well as the Batch Normalization layer are added, which are used to improve the accuracy of the feature extraction network; Second, a feature fusion module is designed to fuse adjacent feature maps to improve the detection effect by integrating contextual information; Third, the SE attention mechanism is introduced to give channel weights adaptively and enhance the useful feature channels; Finally, the object detection analysis experiments are conducted on the PASCAL VOC2012 dataset. The experimental results show that the improved SSD algorithm in this paper is able to achieve an mean average precision of 72.7% in the data set, which is 2.1% better than the original SSD-VGG16 and greatly improves the object detection effect.**

Keywords-Object Detection; Feature Fusion; Attention Mechanism; SSD Algorithm

I. INTRODUCTION

Object detection has become one of the important research directions in the field of computer vision. Its essence is to find the desired target in a complex background image, give the location information of the target, and judge its class. Object detection technology is widely used in face recognition, medical field, traffic research and other aspects [1].

Object detection based on deep learning can be divided into two groups: two-stage object detection method based on region proposal and one-stage object detection method based on regression [4].In the two-stage series, represented by a series of algorithms of regional convolutional neural network (r-cnn), the task of the first stage is to generate a group of target candidate regions, and then send these candidate regions to the second stage. Then coordinate regression and classification are carried out to gradually realize end-to-end object detection. Although the detection accuracy has been improved, the detection speed is slow due to its large network parameters [6].One-stage series algorithms are represented by single-detector YOLO and single-network multi-scale detector SSD. It discards the stage of extracting candidate regions in the two stage method, and directly obtains the category probability and location of the target, making its network structure simpler. Compared with the two stage object detection method, the detection speed of this method has been improved to a certain extent, but the localization accuracy has decreased,

and there is still a problem that the model parameters are too large [6].

SSD [20] (full name: Single Shot Detection) is a typical algorithm of one-stage object detection method. In order to detect objects of different scales, it uses shallow feature maps to detect small objects, and uses deep feature maps to detect large objects.

The SSD algorithm encapsulates the target location and target prediction in the forward operation, and can directly generate the category probability and position coordinate value of the object. The final detection result can be obtained by only one step detection. Although the detection speed is faster, the positioning accuracy is decreased. SSD algorithm uses shallow high-resolution feature layer. Due to the lack of feature expression ability of this layer, there may be missed detection and false detection when detecting small-scale targets.

In order to avoid the above problems, this paper improves the standard SSD object detection algorithm. The residual network structure is introduced, and the adjacent feature maps are fused. At the same time, the SE attention mechanism experiment is carried out on the PASCAL VOC2012 dataset.

II. SSD ALGORITHM PRINCIPLE

Liu et al. proposed SSD algorithm [20] in 2015, which combines the advantages of Yolo's fast speed and fast rcnn's accurate positioning. The main innovations of SSD algorithm are as follows: (1) extracting feature maps at different scales; (2) using prior boxes with different aspect ratios. These two important improvements enable SSD algorithm to efficiently detect targets with different scales [20].

The network structure of SSD is shown in Figure 1.Its detection framework [20] consists of two sections: feature extraction network and multi-scale feature detection network. The feature extraction network adopts the VGG16 network structure for the preliminary extraction of image features. Since the SSD network only needs to extract features without classification in this part, the fully connection layer in VGG16 is replaced by the convolution layer. The second part is used to detect the predicted feature maps of different scales generated by the feature extraction network. The spatial resolution of the shallow feature map is higher than that of the deep feature map, which can more accurately identify the detailed information such as the edge, contour and texture of the image [18].The deep feature map has a large receptive field and strong information representation ability, but it lacks detailed information compared with the shallow feature map detect smaller sized targets and the deep feature map detect larger targets.

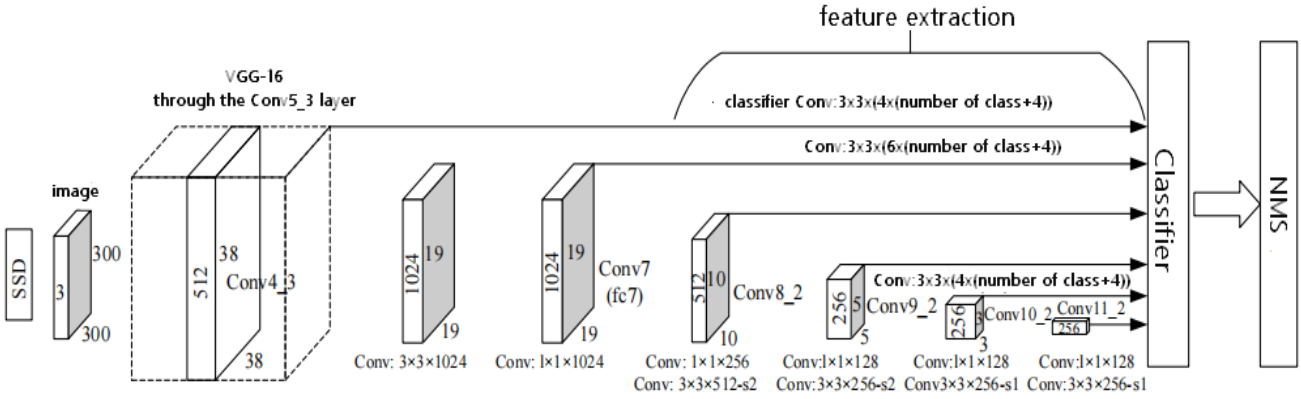


Figure 1. Schematic diagram of SSD network structure

A. Feature extraction network

SSD adopts multi-scale feature detection in object detection. The size of the input image can be 300×300、512×512 [18], etc. The original SSD basic network uses VGG16 as the backbone network. The multi-scale pyramid feature map is generated by adding several convolution layers with gradually reduced size to the basic network: First, the FC7 layer in VGG16 is replaced by the convolution layer Conv7, and all dropout layers and FC8 layers are removed; Secondly, the feature layers Conv8, Conv9, Conv10 and Conv11 are added. The SSD network uses the Conv4_3 layer in vgg16 as the first prediction feature map, and the feature maps obtained from the Conv7, Conv8_2, Conv9_2, Conv10_2, and Conv11_2 layers as the subsequent prediction feature maps. So, a total of 6 prediction feature maps with different scales are obtained. Taking the input image size of 300×300 as an example, the size of the six predicted feature maps obtained by the feature extraction network are 38×38, 19×19, 10×10, 5×5, 3×3, 1×1.

B. Multi-scale feature detection network

For the six predicted feature maps, each cell of each predicted feature map will extract k = 4 to k = 6 default boxes according to different aspect ratios, and finally obtain 8732 prior boxes. If the size of a feature map is H×W, it means that H×W×k default boxes will be generated on this feature map. The k values on the six predicted feature maps are: 4, 6, 6, 6, 4, 4.The size and number of default boxes generated by each feature map are shown in Table Ⅰ.

TABLE I. DEFAULT DIMENSION AND QUANTITY OF FEATURE MAP

Feature Map	Width and height of the feature map	Default boxes size	Number of default boxes
Feature Map1	38 x 38	$21\{1/2,1,2\};\sqrt{21 \text{ x } 45}\ \{1\}$	38 x 38 x 4
Feature Map2	19 x 19	$45\{1/3,1/2,1,2,3\};\sqrt{45 \text{ x } 99}\ \{1\}$	19 x 19 x 6
Feature Map3	10 x 10	$99\{1/3,1/2,1,2,3\};\sqrt{99 \text{ x } 153}\ \{1\}$	10 x 10 x 6
Feature Map4	5 x 5	$153\{1/3,1/2,1,2,3\};\sqrt{153 \text{ x } 207}\ \{1\}$	5 x 5 x 6
Feature Map5	3 x 3	$207\{1/2,1,2\};\sqrt{207 \text{ x } 261}\ \{1\}$	3 x 3 x 4
Feature Map6	1 x 1	$261\{1/2,1,2\};\sqrt{261 \text{ x } 315}\ \{1\}$	1 x 1 x 4

After obtaining all default boxes, we will calculate the scores of c categories and 4 coordinate offsets (boundary box regression parameters) for each default box: For the M×N, P-channel predicted feature map, two 3×3 convolution kernels of P channels are used to generate probability scores and coordinate offsets of corresponding default boxes. Let s be the number of all default boxes on these six prediction feature maps, then the number of convolution kernels required for calculating the category is s × c, and the number of convolution kernels required to calculate the coordinate offset is s ×4.

The loss function of multi-scale feature detection network is divided into two parts: category loss and position offset loss. The calculation formula of the target loss function is shown in (1). Among them, N is the number of matched positive samples, α is 1, $L_{conf}(x,c)$ is the predicted category loss, $L_{\text{loc}}(x,l,g)$ is the position offset loss.

$$L(x,c,l,g)=\frac{1}{N}(L_{conf}(x,c)+\alpha L_{loc}(x,l,g)) \quad (1)$$

For the predicted category loss, the calculation formula is shown in (2).

$$L_{\text{conf}}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} \log(\hat{c}_i^{p}) - \sum_{i\in Neg} \log(\hat{c}_i^{0})$$
$$\text{where } \hat{c}_i^{p} = \frac{\exp(c_i^{p})}{\sum_{p}\exp(\hat{c}_i^{p})} \quad (2)$$

Among them, $-\sum_{i\in Pos}^{N} x_{ij}^{p} \log(\hat{c}_i^{p})$ is the category loss of positive samples and $-\sum_{i\in Neg}\log(\hat{c}_i^{0})$ is the category loss of negative samples;

$\hat{c}_i^{p}$ is the category probability p of predicting the GT box corresponding to the i-th default box.

x_{ij}^{p} ={0,1} is the j-th GT box (category is p) matched by the i-th default box.

$\hat{c}_i^{0}$ is the probability that the i-th default box is predicted to be the background.

The equation of positional offset loss $L_{\text{loc}}(x,l,g)$ is shown in (3). Among them, l_i^m is the predicted regression parameters (center coordinate x, center coordinate y, width w, height h) corresponding to the i-th positive sample; g_j^m is the coordinate of the j-th GT box matched to the positive sample i; d_i^m is the coordinate of the i-th default box; $\hat{g}_j^m$ is the regression parameter calculated from the coordinates of the i-th default box and the coordinates of the j-th GT box matched to it.

$$L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k} smooth_{l1}(l_i^m - \hat{g}_i^m)$$
where
$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h,\ \ \hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h$$
$$\hat{g}_j^{w} = \log(\frac{g_j^{w}}{d_i^{w}}),\ \ \hat{g}_j^{h} = \log(\frac{g_j^{h}}{d_i^{h}}) \quad (3)$$
$$smooth_{l1}(x) = \begin{cases} 0.5x^2, & |x|<1 \\ |x|-0.5, & other \end{cases}$$

III. IMPROVED SSD NETWORK STRUCTURE

The improvements in this paper based on the SSD algorithm include:

1) In this paper, ResNet50 network is used as a feature extraction network. First, the step of the first Residual structure of the Conv4 layer in the ResNet50 network is changed from 2 to 1; second, the structures of Conv5, Avg Pool, FC and Softmax behind the Conv4 layer are deleted, so the first predicted feature map is obtained; finally, five downsampling operations are added after the first predicted feature map. Among them, the Batch Normalization layer is added after each convolution operation; in addition, the bias parameter terms in the convolution operation and the weight decay of the parameters of the Batch Normalization layer are removed. The network structure is shown in Figure 2.

2) A total of 6 predicted feature maps are obtained from the operation of 1). To enhance the semantic information of the feature maps, the shallow feature maps are fused with the deep feature maps to obtain the new prediction feature maps. The new prediction feature maps are input to the network for subsequent operations.

3) Introducing the SE attention mechanism. After the fused predicted feature maps are input to the SE attention module, the SE attention mechanism obtains the importance of each channel in the feature maps. Based on this importance level, the SE attention mechanism assigns a weight value to each feature channel so that the feature channels that are useful for the current task are enhanced and the feature channels that are not useful for the current task are suppressed.

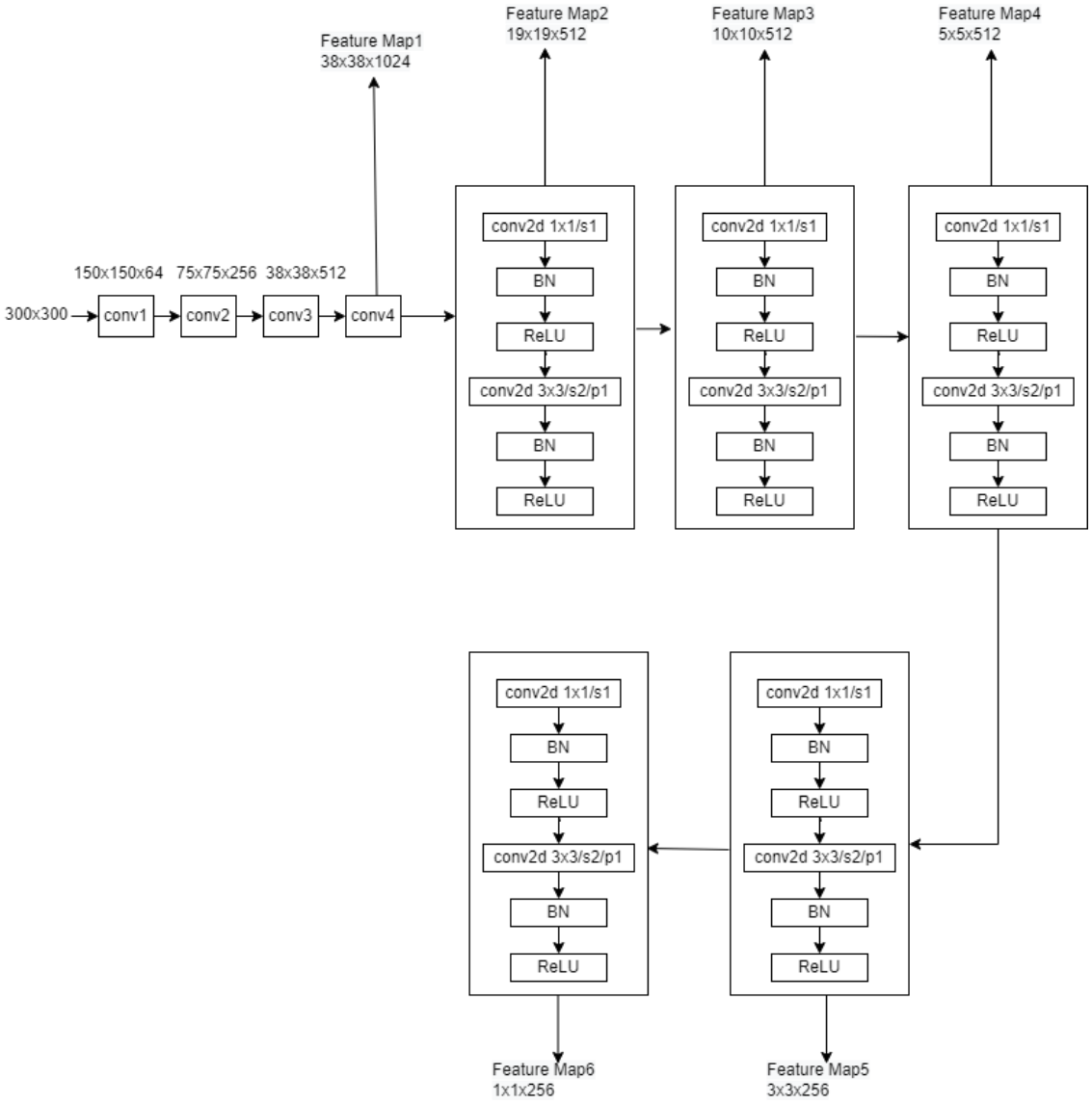


Figure 2. Schematic diagram of the base network with improved SSD

A. The Improvement of Backbone Network

In this paper, the backbone in the SSD network is replaced by the VGG16 network with the ResNet50 network. In the object detection algorithm, the selection of the classification network has a significant impact on the performance of the algorithm. Removing the fully connected layer and loss layer from the classification network will get the basic network part of the object detection algorithm.VGG16 consists of 13 convolutional layers, 3 fully connected layers and 5 pooling layers. The outstanding feature of vgg16 is its simple structure, which is easier to build by stacking several convolutional and pooling layers. However, the disadvantage of VGG16 is that the number of layers is shallow and the feature extraction is not sufficient[8].The ResNet50 network consists of a series of Residual structures and uses Batch Normalization to accelerate the training; and the ResNet50 network is much smaller than VGG16, and its speed and accuracy are superior to VGG16. Therefore, in this paper, ResNet50 is chosen to replace the VGG16 network in the SSD network.

B. Feature Fusion

The shallow feature maps are large in size and contain sufficient detail information, but too little semantic information. The deep feature maps have large receptive fields and contain sufficient semantic information, but detail information is gradually lost as the receptive field increases [3]. Therefore, fusing features of different scales is an important means to improve the performance of object detection. In this paper, bilinear interpolation is used to achieve upsampling, which increases the image resolution and retains more feature information; the concatenate method is used to stitch the shallow feature map with the deep feature map in the depth direction. The feature fusion method is shown in Figure 3.

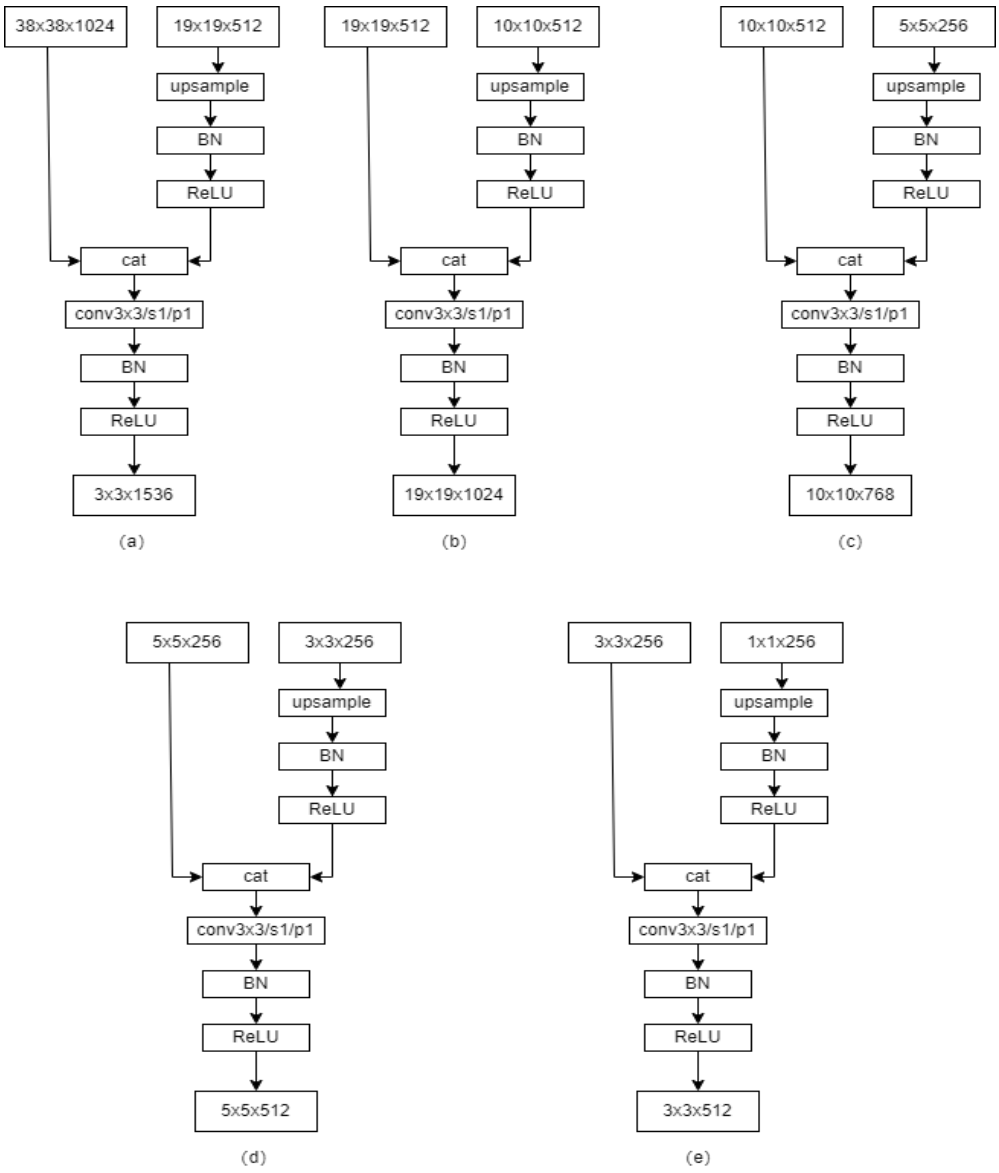


Figure 3. Schematic diagram of the fusion of shallow features and deep features

Taking the fusion of Conv4 feature layer and Conv7 feature layer as an example, as shown in Figure 3(a), the specific fusion is as follows:

1) First, the Conv7 feature layer is first upsampled by bilinear interpolation, Batch Normalization and linear activation (ReLU), so that the width and height of the feature map of this

layer are the same as the width and height of the feature map of the Conv4 layer. Then, the feature map of size 38×38×512 is obtained.

2) Second, the feature map obtained from the Conv7 layer is concatenated with the feature map from the Conv4 layer to obtain a feature map of size 38×38×1536.

3) Finally, the feature map of size 38×38×1536 is subjected to a 3×3 convolution, batch normalization and linear activation (ReLU) to obtain the final first predicted feature map.

In Figure 3(b), the feature maps of the Conv7 layer are fused with the feature maps of the Conv8_2 layer, and the obtained feature maps are used to replace the feature maps of the Conv7 layer.

In Figure 3(c), the feature maps of the Conv8_2 layer are fused with the feature maps of the Conv9_2 layer, and the feature maps of the Conv8_2 layer are replaced with the resulting feature maps.

In Figure 3(d), the feature maps of the Conv9_2 layer are fused with the feature maps of the Conv10_2 layer, and the feature maps of the Conv9_2 layer are replaced with the obtained feature maps.

In Figure 3(e), the feature maps of the Conv10_2 layer are fused with the feature maps of the Conv11_2 layer, and the feature maps of the Conv10_2 layer are replaced with the obtained feature maps.

The operation flow of their fusion is similar to Figure 3(a), and the final size of the new predicted feature layers are obtained as:

$38\times38\times1536, 19\times19\times1024, 10\times10\times768, 5\times5\times512, 3\times3\times512, 1\times1\times256$.

C. SE attention mechanism

SENet emerged to solve the loss problem caused by the different importance of different channels of the feature map in the convolutional pooling process. In the traditional convolutional pooling process, each channel of the feature map is equally important by default. In real-world problems, the importance of different channels varies. SENet can adaptively recalibrate the channel feature responses by explicitly modeling the interdependencies between channels. The role of SENet is to obtain the weights of each channel of the incoming feature map, allowing different channels to have different effects on the task results with different weights. So, the use of SENet allows the network to focus on the channels that need to play the most role in the detection task.

The process of SEblock is divided into two steps: Squeeze and Excitation.

(1) Squeeze: Obtaining the global compressed feature volume of the current feature map by performing Global Average Pooling on the feature map layer.

(2) Excitation: The weights of each channel in the feature map are obtained by a two-layer fully connected bottleneck structure, and the weighted feature map is used as the input of the next layer of the network.

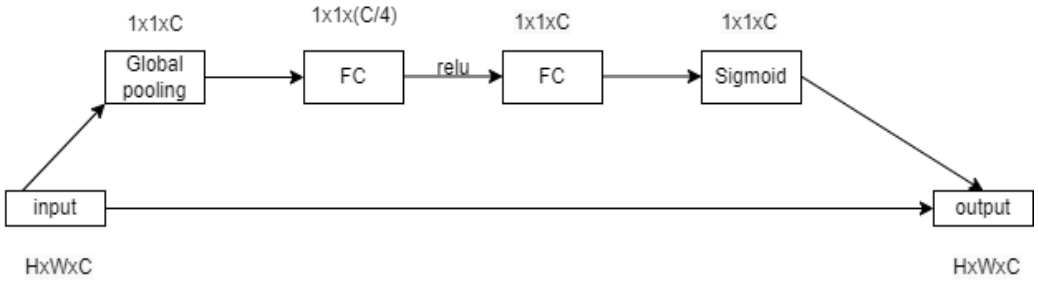


Figure 4. Schematic diagram of SE module

The four steps of the SE attention mechanism are illustrated in Figure 4, as follows:

(1) Performing global average pooling of the input feature maps.

(2) Performing two full connections, the first with a smaller number of fully connected neurons and the second with the same number of fully connected neurons as the number of channels of the input feature map.

(3) After completing two full connections, the weights (between 0 and 1) of each channel in the input feature map are obtained by performing another Sigmoid to fix the value between 0 and 1.

(4) After obtaining this weight value, multiply this weight value by the input feature map.

The feature map contains a large number of channels, and the judgment of the network varies from channel to channel. Some channels contain rich information and some channels hardly play a role. The purpose of introducing the SE attention mechanism is to make full use of the feature channels that contain important information. The SE attention mechanism is added to the back of the six predicted feature maps obtained after feature fusion, so that it adaptively assigns weights to each channel in the feature maps, strengthening the useful feature channels while suppressing the useless ones. The SSD model with the added SE attention mechanism is shown in Figure 5.

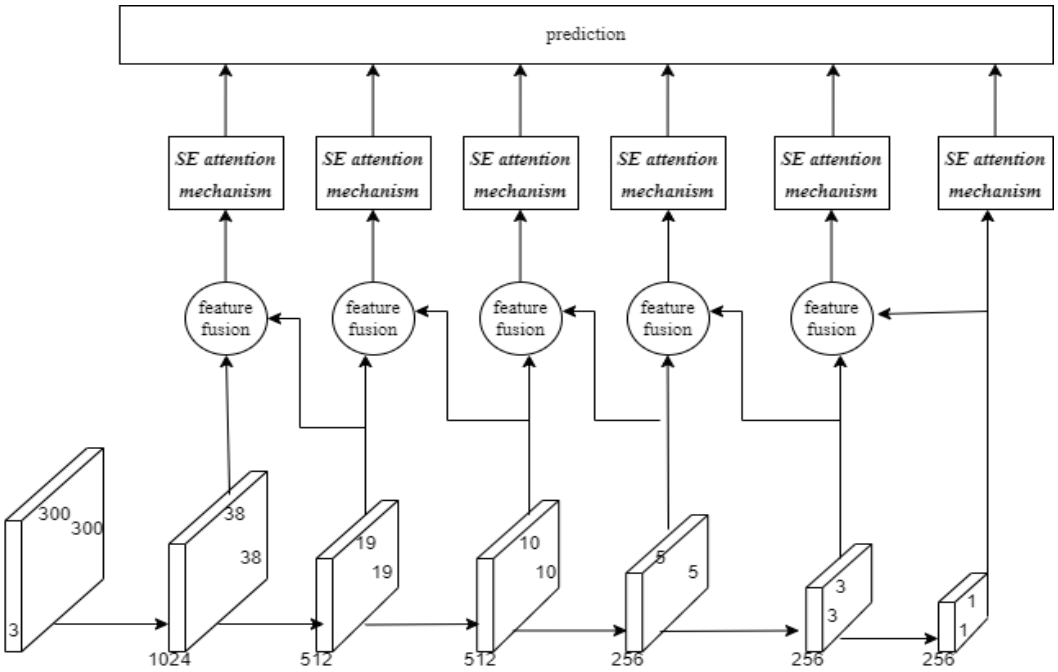


Figure 5. Schematic diagram of the network model with improved SSD

IV. EXPERIMENTAL PROCEDURE AND ANALYSIS OF RESULTS

A. Preparation of dataset and development environment

This experiment uses the PASCAL VOC2012 dataset, which contains 11530 images and 27450 targets labeled. There are 20 categories of targets to be recognized in this dataset, which are divided into four main categories: Vehicles, Household, Animals, and People; Vehicles class include Aero plane, Bicycle, Boat, Bus, Car, Motorbike, and Train; Household class contains: Bottle, Chair, Dining table, Potted plant, Sofa, TV/Monitor; Animals class contains: Bird, Cat, Cow, Dog, Horse, Sheep. The images in the dataset are annotated with corresponding XML files for the location and class of the target [4].

The environment used for the experiments is shown in Table Ⅱ.

TABLE II. ENVIRONMENT CONFIGURATION TABLE

Hardware	**Processor** **Video Cards**	**Intel(R)Core(TM) i7-6500U** **GeForce_RTX_2080_Ti**
Software	Operating System	windows10
	Deep Learning Framework	pytorch-gpu
	Compiler Language	python
	Compilers	pycharm

B. Evaluation Indicators

To comprehensively evaluate the accuracy of the SSD algorithm in detecting targets, this paper chooses to use the Mean Average Precision (mAP) as the evaluation criterion. MAP represents the average of all Average Precision (AP), and each Average Precision(AP) is measured using the intersection over Union (IOU). Samples with IOU above the threshold are positive samples, and samples below the threshold are negative samples. The calculation of AP requires Precision, Recall, as shown in (4):

$$\begin{aligned} \text{Presion} &= \text{TP}/(\text{TP}+\text{FP}) \\ \text{Recall} &= \text{TP}/(\text{TP}+\text{FN}) \\ AP &= \int_0^1 p(r)dr \end{aligned} \tag{4}$$

Among them,TP is the positive sample with positive prediction, FP is the negative sample with positive prediction, FN is the positive sample with negative prediction, and p(r) is the precise-recall curve.

C. Setting of training parameters

TABLE III. TRAINING PARAMETERS SETTING TABLE

Parameter	**Value**
learning rate	0.0005
momentum	0.9
weight_decay	0.0005
batch size	16
epoch	50
step_size	5

D. Experimental results and analysis

In the experiments, transfer learning strategy is used to reduce the training difficulty and improve the detection results. The improved SSD algorithm in this paper is trained on the ResNet50 pre-training model, and its loss function remains the same as that of the original SSD-VGG16 algorithm. After the training, the improved algorithm in this paper produces the loss curve shown in Figure 6. From this figure, it can be seen that the algorithm in this paper starts to converge when the iteration reaches the 25th epoch. The final training loss value is 2.72.

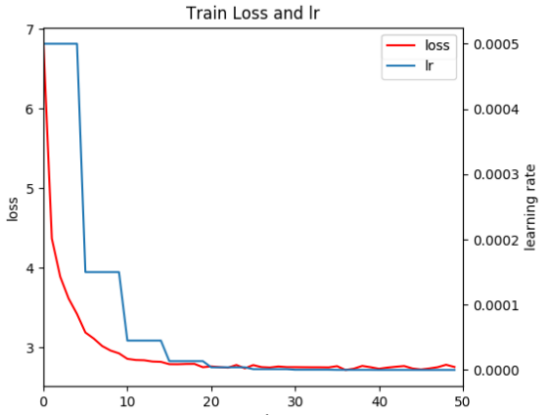


Figure 6. Loss curve of improved SSD algorithm

The distribution of mAP values is shown in Figure 7. It can be seen that the mAP value reaches a maximum of 72.67%, which is an increase of 2.1% over the original SSD-VGG16 network.

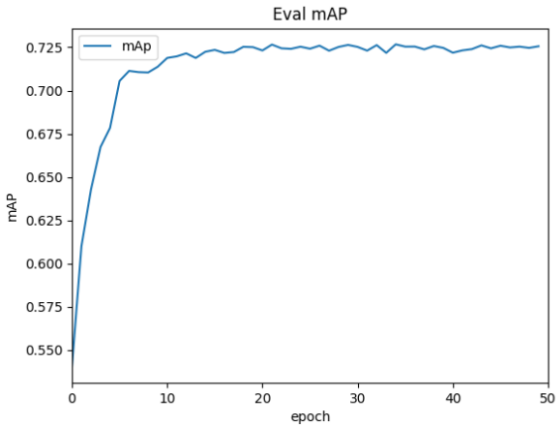


Figure 7. mAP graph of improved SSD algorithm

In this paper, the SSD-VGG16 algorithm, the SSD-ResNet50 algorithm and the improved algorithm of this paper are respectively trained and validated on the PASCAL VOC2012 dataset. And

we calculates the mAP of each algorithm to detect the target. From Table Ⅳ, we can see that the mAP of the improved algorithm in this paper improves from 70.6% to 72.7% compared with the original SSD-VGG16 algorithm, thus the detection effect of this paper's algorithm is better than the original SSD algorithm.

TABLE IV. COMPARISON OF DETECTION RESULTS OF DIFFERENT ALGORITHMS ON THE PASCAL VOC2012 DATASET

Algorithm	mAP
SSD-VGG16	70.6%
SSD-ResNet50	72.1%
The improved algorithm	72.7%

V. CONCLUSION

In order to improve the detection accuracy of SSD algorithm for object detection, this paper improves the network structure of SSD object detection algorithm, and introduces multi-layer feature fusion module and channel attention mechanism. Multi-layer feature fusion at different scales based on the ResNet50 feature network increases the semantic information of the original feature map. At the same time, the SE attention mechanism is added to enhance the focus on feature channels for each layer of the network. The comparison experiments on the PASCAL VOC2012 dataset show that the improved SSD algorithm in this paper has a 2.1% higher mean average precision and better detection capability compared to other commonly used algorithms.

REFERENCES

[1] Li Weiqiang, Wang Dong, Ning Zhengtong, Lu Mingliang, Qin Pengfei. A review of fruit target detection algorithms under computer vision [J]. Computers and Modernization, 2022(06):87-95.

[2] Lv Lu, Cheng Hu, Zhu Hongtai, Dai Nianshu. A review of target detection research and application based on deep learning [J]. Electronics and Mounting, 2022, 22(01):72-80.DOI:10.16257/j.cnki.1681-1070.2022.0114.

[3] Deng Quan, Lin Xingxing. Improved SSD-based algorithm for marine life detection [J]. Computer Technology and Development, 2022, 32(04):51-56.

[4] Xie F, Zhu D.B.. A review of deep learning target detection methods [J]. Computer Systems Applications, 2022, 31(02):1-12. doi:10.15888/j.cnki.csa.008303.

[5] Peng Hongxing, Li Jing, Xu Huiming, Chen Hu, Xing Zheng, He Huijun, Xiong Juntao. Litchi detection based on multiple feature enhancement and feature fusion SSD [J]. Journal of Agricultural Engineering, 2022, 38(04):169-177.

[6] Jia Kexin, Ma Zhenghua,Zhu Rong,Li Yonggang. Attention mechanism to improve lightweight SSD model for sea surface small target detection [J]. Chinese Journal of Graphics, 2022, 27(04):1161-1175.

[7] Zheng Qiumei, Xu Linkang, Wang Fenghua, Lin Chao. Pyramid scene resolution network based on improved self-attention mechanism[J/OL]. Computer Engineering:1-9[2022-07-08]. DOI:10.19678/j.issn.1000-3428.0063652.

[8] Huang Yichuan, Li Lianghai, Ma Jijun, Cui Huimin. Research on aerial photography target detection algorithm based on improved SSD algorithm [J]. Telemetry and Remote Control, 2022, 43(03):79-85.

[9] Gao Na, Wu Qing, Zhang Man-ho. Multi-scale feature enhancement algorithm for SSD target detection [J]. Journal of Hebei University of Technology, 2022, 51(02):23-30.

[10] Guo Jianzhong, Yu Tengfei, Cui Yuding, Zhou Xinglin. Research on improved SSD-based vehicle small target detection algorithm [J]. Computer Technology and Development, 2022, 32(03):1-7.

[11] Yin, Meng-Yuan. Research on SSD target detection algorithm based on attention mechanism[C]. Anhui University of Technology, 2021.

[12] Sun Peng. Research on small target detection based on improved SSD model[C]. Nanjing University of Posts and Telecommunications, 2021.

[13] Yao Guanghua, Wu Xuncheng, Zhang Xuexiang, and Squire Jun. A small target detection method incorporating contextual information features [J]. Computer and Digital Engineering, 2022, 50(05):1018-1022.

[14] Yang Haojie, Wang Lu, Yang Shengwei. A feature fusion-based road target detection method [J]. Journal of Changsha University, 2022, 36(02):1-6.

[15] Li Kequan, Chen Yan, Liu Jia Chen, Mou Xiangwei. A review of deep learning-based target detection algorithms [J/OL]. Computer Engineering:1-17[2022-07-09]. DOI:10.19678/j.issn.1000-3428.0062725.

[16] Liu X. Research on improved SSD-based target detection algorithm [D]. Xiangtan University, 2021. DOI:10.27426/d.cnki.gxtdu.2021.002054.

[17] Wang Yanni, Yu Lixian. Attention and multi-scale effective fusion of SSD target detection algorithm [J]. Computer Science and Exploration, 2022, 16(02):438-447.

[18] He Jingyuan, Xie, Shenglong, Tian Yuan, Tian Qinqin. Multi-scale feature fusion for target detection algorithm [J]. Henan Science, 2021, 39(07):1045-1051.

[19] Wu T-C, Wang X-T, Cai Y-J, Jing Y-B, Chen C-Y. Lightweight SSD target detection method based on feature fusion [J]. Liquid Crystal and Display, 2021, 36(10):1437-1444.

[20] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]// European Conference on Computer Vision. Springer, Cham, 2016: 21-37.

Research on Topic Fusion Graph Convolution Network News Text Classification Algorithm

Li Liu
School of Computer Science and Technology
Xi'an Technological University
Xi'an, China
E-mail: Lily_Liu20@163.com

Yongyong Sun
School of Computer Science and Technology
Xi'an Technological University
Xi'an, China
E-mail: yongsunjd@126.com

***Abstract*—In the face of a large amount of news text information, how to make a reasonable classification of news text is a hot issue of modern scholars. To solve the problem that only word co-occurrence was considered in the Text Graph Convolutional Network (Text-GCN) method to build a graph model, a news text classification algorithm which fuses themes and is based on Graph Convolution Network, is presented. Firstly, the LDA topic model is used to process the corpus to obtain the distribution of themes of the corpus. Secondly, a graph model is built to construct a global map by using the related topic words and their subject distribution in each article. Finally, the text graph is input into the Graph Convolution Network layers to compute the learning representation of combining feature in order to complete the text classification task. The experimental results show that this method can effectively realize the word level interaction of information in text. In the experiment on Chinese and English datasets, adding theme information improves the accuracy by 1% compared with the Text-GCN method.**

Keywords-Text Graph Convolutional Network; Graph Convolutional Network; Latent Dirichlet Allocation Topic Model(LDA topic model); Graph Model

I. INTRODUCTION

With the rapid development of the Internet and its related fields, a large amount of data is continuously imported into the network, of which text data accounts for the vast majority. In the environment of information explosion, obtaining news information through the network has become an important way for people to understand the world and the era of science and technology. How to effectively analyze and use these data has become an urgent problem to be solved. Text classification is the work of mapping a text to one or more categories by analyzing the inherent characteristic information of the text. It is the basis of intelligent text reading. Text classification is an indispensable part of analyzing and processing a large number of text data, which is very challenging. It can serve a variety of downstream tasks, such as Machine Translation, Text Summarization, Recommendation Systems, etc., and is the focus of academic and industrial circles. Therefore, to a certain extent, accurate text classification can not only facilitate its downstream tasks, but also shorten the time for users to search text information and avoid information overload.

DOI: 10.2478/ijanmc-2022-0030

In the field of natural language processing (NLP), scholars have been trying to take the characters, words and sentences contained in documents as the starting point, and use their own features and associated features to integrate learning documents to achieve the final classification task. Text classification tasks can be divided into two categories: manual classification and automatic classification [1]. Manual classification methods often use professionals to interpret the content, compile a set of language rules after translation, and then classify according to this rule. This method only relies on manual processing, and there are great challenges in the face of a large number of data sets. It usually costs a lot of time and expensive cost, and the classification accuracy is easily affected by human factors. Automatic classification methods can be divided into machine learning based and depth learning based methods. Although the method based machine learning relies on a variety of calculation methods such as Bayesian, Laplace, Laplace's derivation of least square method and Markov chain to process information and can bring significant performance improvement, it is still limited by the characteristics of manual design. With the continuous development of deep learning technology, neural network provides a new way to solve complex problems in various fields, such as intelligently extracting text content and constructing text feature learning documents by using Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) or Graph Neural Network (GNN).

In recent years, further research on GNN and graph embedding has also attracted extensive attention of scholars [2-3]. Generally, in the task of text classification, the feature engineering of text information often uses sequence data for expression, and how to effectively express text information with graph data structure, that is, the representation method of text is an important basis for the further development of graph neural network and graph embedding in the direction of NLP. This paper constructs a large heterogeneous graph, which can model the self-information and auxiliary information of the text, and connect the explicit information and potential semantic information of the text through co-occurrence. At the same time, a text graph convolution classification algorithm is proposed. This method can effectively capture the three kinds of information of documents, topics and words in heterogeneous information networks, as well as the relationship between them, enhance the semantic information contained in document nodes, and effectively alleviate the semantic deficiency caused by the graphic representation of text.

II. RELATED WORK

The text classification problem is divided into two parts: feature engineering and classifier. Feature engineering transforms text into text representation data. The traditional method is mainly manual processing, and the deep learning method takes word embedding as the core. The classifier uses Support Vector Machine (SVM), decision tree or neural network to classify the text representation data with features. The text classification method based on deep learning takes the pre-labeled samples as training data, constructs a learning model, learns and infers the internal relationship between the text and its labels, and classifies and predicts the new text set according to the learned correlation [1]. It can complete the text representation in large-scale text classification tasks. Some recent studies have shown that the effectiveness of word embedding can largely affect the accuracy of deep learning classification methods [4].

A. *Text representation of sequence information*

Traditional feature engineering uses word2vec and Bag-of-Word methods to realize text representation. However, the words in the text are not simply listed, but have certain rules. Kim [5] proposed TextCNN (Text Convolutional Neural Networks) model, which maps the text into a vector, then uses multiple filters to capture the local semantic information in the text as features, and finally obtains the probability distribution of classification labels through the full connection layer. Lai [6] et al. proposed Recursive Convolutional Neural Networks (RCNN), considering the context representation of words to obtain the probability representation of labels. This kind of text representation gives priority to the order or local information of the text after text data preprocessing. Although they can well capture the semantic and grammatical information in continuous sequences, they do not analyze the internal structure of the text.

B. *Text representation of graphic information*

In the field of natural language, text has not only the sequential expression of context, but also the internal graph structure. For example, syntactic and semantic parsing trees define the syntactic and semantic relationships between words in sentences. TextRank [7], the earliest one based on graph model, proposed to represent natural language text as a graph $G(V,E)$, V represents a group of nodes, and E represents a group of edges between nodes. Nodes can represent various types of text units, such as words, phrases or sentences in a text. Edges represent different types of relationships between any node, such as lexical or semantic relationships, context overlap relationships, and so on. Under the influence of the TextRank model, scholars take words, phrases and sentences in the document as nodes to build different graph models to effectively express text features. Yao [8] et al. proposed the Text-GCN (Text Graph Convolutional Networks) model, which is based on the whole corpus. By constructing text maps and using GCN to jointly learn the maps, words and documents can be embedded, and then the classification results can be obtained through *softmax* function. The author uses words and documents as nodes to build a heterogeneous graph. Considering that global word co-occurrence will carry discontinuous and long-distance semantic information, the author calculates edge weights through TF-IDF, and uses word frequency and document frequency to build edges between word nodes and document nodes. The edges between two word nodes are represented by word co-occurrence. Text-GCN model transforms text classification problem into node classification problem, which improves the accuracy of classification results. However, this method ignores the relationship between context words in the text, and does not consider the fine-grained text level word interaction. Instead, all words are regarded as independent individuals without upper and lower attachments, and some document information is lost. In order to solve this problem, Zhang [9] et al. proposed to express inductive words through graph neural network, that is, texting. The author believes that each text is a separate graph, in which text level word interaction can be learned, and graph neural network is added to train the text to describe the detailed word-word relationship. Liu [10] et al. proposed TensorGCN (Tensor Graph Convolutional Networks) to describe multiple semantic, grammatical and contextual information by constructing text graph tensors. Because the above method considers the context, the model can learn the emotional preference of the text by capturing the semantic emotional information of the context, and the accuracy of the emotional text classification task has been improved. However,

when considering fine-grained word-word information, these methods construct multiple graphs for each text to cover the context information contained in the text, and then use multi-layer convolution calculation to realize text classification. The calculation and storage of multiple information graphs add a lot of storage and computational burden to the algorithm.

To sum up, in order to improve the classification effect, this paper proposes a topic fusion graph convolution neural network text classification algorithm. In this method, all words in the corpus are considered. By inserting topic information, LDA Topic Heterogeneous Graph (LDA-THG) is generated to further improve the accuracy of text classification.

III. METHOD

In this paper, the text classification model based on LDA-THG is proposed. After preprocessing the text documents, the LDA Topic model (Latent Dirichlet Allocation Topic Model) is used to conduct topic training on the global corpus. Then, the probability distribution obtained by training is combined with the self-information of each text to conduct text modeling. This method constructs semantic association information among words, topics and documents, and uses graph convolutional network to process the embedded information of document nodes and their neighborhood nodes, so as to alleviate the problem of insufficient semantic information when using graph as text representation in text classification. The framework of text classification algorithm model based on LDA-THG graph convolution is shown in Figure 1.

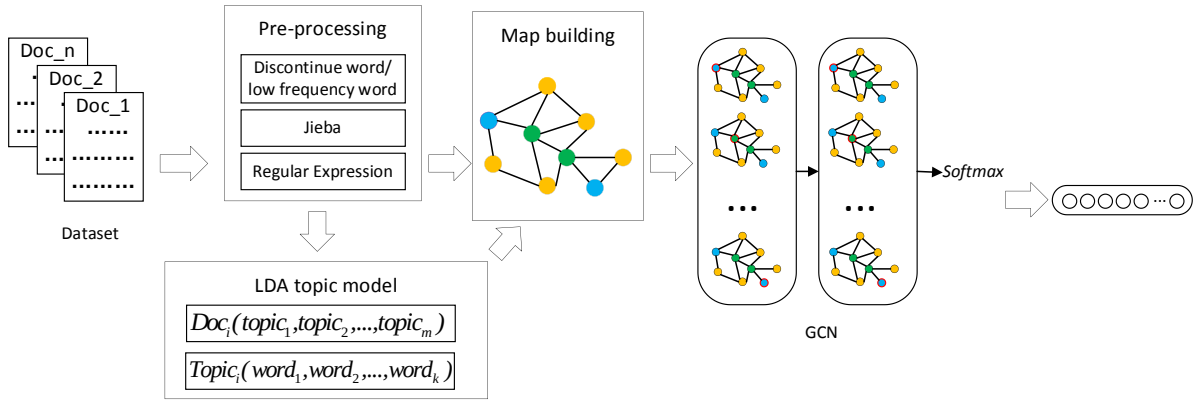


Figure 1. Framework of Graph Convolution Text Categorization Algorithm Based on LDA-THG

A. Graph Convolutional Network

Graph Convolutional Network [11] is a multi-layer neural network that transforms the traditional convolutional neural network that processes regular spatial structure data into a multi-layer neural network that processes graph data. Generally, when processing graph G, it is assumed that each node in the graph can form a connection with itself, that is $\forall v \in V$ and $(v,v) \in E$. Let $X = (x_1^T, x_2^T, x_3^T, ..., x_n^T)$, $X \in \mathbb{R}^{n \times m}$ be a matrix containing all nodes and their characteristics, where m is the dimension of the eigenvector, and each row $x_v \in \mathbb{R}^m$ is the eigenvector of node v. The adjacency matrix $A \in \mathbb{R}^{n \times n}$ and degree matrix $D \in \mathbb{R}^{n \times n}$ of graph G

are introduced, where $D_{ii} = \sum_j A_{ij}$. Due to self-connection and self-circulation, the diagonal element of adjacency matrix A is set to one, where $A_{ii} = 1$. When GCN only passes through one-layer of convolution, it can capture the nearest neighbor information of nodes. When multiple GCN layers are stacked, it will integrate more information.

For one-layer GCN, the new k-dimensional node characteristic matrix $H^{(1)} \in \mathbb{R}^{n \times k}$ is calculated as:

$$H^{(1)} = \sigma(AXW^{(0)}) \quad (1)$$

Where, $A = D^{-1/2} A D^{-1/2}$ is the normalized symmetric adjacency matrix, $W^{(0)} \in \mathbb{R}^{m \times k}$ is the weight matrix, and $H^{(0)} = X$. σ is the activation function, such as ReLU. Similarly, higher neighborhood information can be merged by stacking more GCN layers, namely:

$$H^{(l+1)} = \sigma(AL^{(l)}W^{(l)}) \quad (2)$$

Where l represents the number of layers of the graph, $H^{(l+1)}$ is the output characteristic matrix of layer l, and $W^{(l)}$ the weight matrix of layer l.

B. Construction of LDA-THG

In graph-based convolution algorithm, LDA-THG constructs a large heterogeneous graph. It is a fusion of topics and contains three types of nodes-words, documents and topics. The structure of LDA-THG diagram is shown in Figure 2. This method can explicitly model and realize topic and global word co-occurrence, making it easier to use GCN for joint learning reasoning.

In Figure 2, Doc_n represents document node, Topic_m represents the topic node extracted from all documents in the corpus through LDA model, word_k is the node of the only word in the corpus. The curves in the figure are the edges of figure G, which connect document nodes and word nodes, word nodes and topic nodes, word nodes and word nodes, document nodes and topic nodes, respectively. The number of nodes is $|V| = N + K + M$ in the text graph G, where N is the number of all documents in the corpus, K is the number of all non repeated words in the corpus, M is the number of topics extracted through LDA model.

In this paper, one-hot encoding is used to initially represent the features, that is, the feature matrix of the graph is simply set as a unit matrix, and each node in the graph is represented as a one-hot vector as the input of GCN. For the co-occurrence information of global words, a fixed size sliding window is used to collect the co-occurrence statistical information of all documents in the corpus. For topic information, the whole corpus is processed by LDA topic model to obtain topic distribution θ and word distribution φ. The distribution process of topics and words obtained by the LDA model is shown in Figure 3.

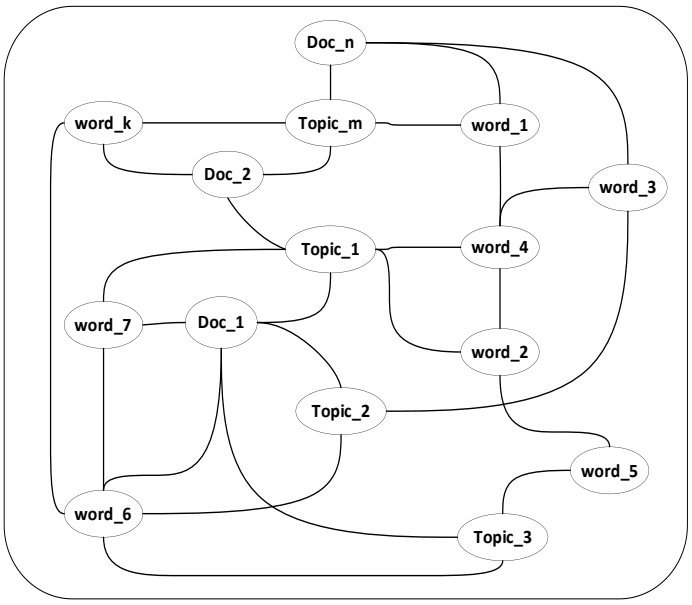


Figure 2. The structure of LDA-THG

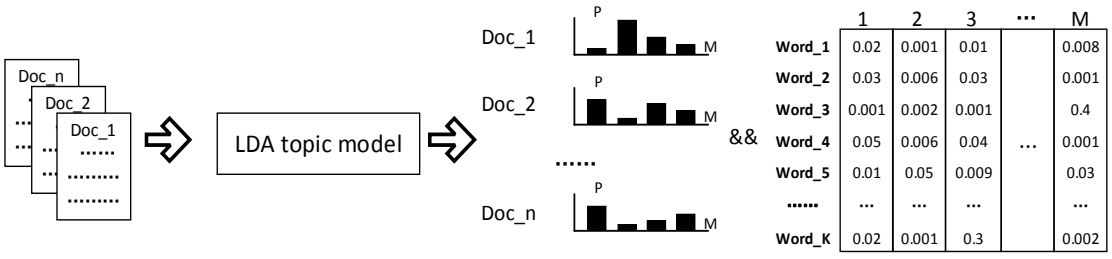


Figure 3. The process diagram of topic and word distribution obtained by LDA model

According to the word frequency in the document, the coexistence of words in the whole corpus, and the topic distribution and word distribution obtained by LDA processing the whole corpus, the edges between nodes are constructed. The weight of the edges between two nodes i and j are defined as:

$$A_{ij}=\begin{cases} PMI(i,j) & , \quad i \text{ and } j \text{ are words} \\ DT(i,j)=\theta_{ij} & , \quad i \text{ is doc}, \ j \text{ is topic} \\ TF\_IDF(i,j) & , \quad i \text{ is doc}, \ j \text{ is word} \\ TW(i,j)=\varphi_{ij} & , \quad i \text{ is topic}, \ j \text{ is word} \\ 1 & , \quad i=j \\ 0 & , \quad otherwise \end{cases} \tag{3}$$

In formula (3), PMI[8] ratio is point level mutual information, which represents the association measure of a popular word and can be used to calculate the association weight between word node i and word node j. And the value of PMI is always greater than zero. $DT(i,j)$ is the weight of the edge between the document node and the topic node, θ_{ij} is the probability that the topic of the i text is j. $TF\_IDF(i,j)$ is the weight of the edge between the document node and the

word node. It is the product of the word frequency of word j in document i and the inverse document frequency of word j. $TW(i,j)$ is the weight of the edge between the topic node and the word node, φ_{ij} is the probability that the content described by word j is related to the i-th topic.

C. Model training

After the heterogeneous graph is successfully constructed, the text graph is input into the GCN layer. As the input text graph is a heterogeneous graph with three types of nodes and interactive information between nodes, a large feature matrix $X \in \mathbb{R}^{|V|\times|V|}$ is used to represent the LDA-THG, namely:

$$X = X_D \oplus X_T \oplus X_W \tag{4}$$

Among them, $\oplus$ is the connecting symbol of the matrix to splice the three matrices. $X_D \in \mathbb{R}^{N\times|V|}$, $X_D \in \mathbb{R}^{N\times|V|}$, $X_D \in \mathbb{R}^{N\times|V|}$ are the characteristic matrix representing document nodes, topic nodes and word nodes, successively. Each row in these matrices represents the eigenvector of a node, i.e $x_i \in \mathbb{R}^{1\times|V|}$.

In this paper, three-layer GCN is selected in the graph convolution network layer to process the input features and send them into a *softmax* classifier to predict the label of the document. The calculation process is as follows:

$$Z = soft\,max(\,A\sigma(\,A\sigma(\,AXW^{(0)}\,)W^{(1)}\,)W^{(2)}\,) \tag{5}$$

With formula (1) and (2), the weight parameters of the first, second, and third layers are, respectively, $W^{(0)}$, $W^{(1)}$, $W^{(2)}$. They are trained and optimized by the gradient descent method, and $soft\,max(\,x_i\,) = x_i \big/ \sum_i x_i$. The loss function used is the cross-entropy error on all documents, and *L2* regularization term is added to reduce the over fitting of the model. We define D_{train} to represent the set of text indices used for training, C to represent the number of label categories, Y to represent the corresponding tag indicator matrix, Θ to represent model parameters and η to represent regularization factors. The calculation method is as follows:

$$\mathcal{L} = -\sum_{i\in D_{train}} \sum_{j=1}^{C} Y_{ij} \bullet log\, Z_{ij} + \eta \left\| \Theta \right\|_2 \tag{6}$$

In formula (4), $E_{1,2} = A\sigma(\,AXW^{(0)}\,)W^{(1)}$ contains the embedding of first-layer and second-layer documents, words and topics, while $E_3 = ARe\,LU(\,ARe\,LU(\,AXW^{(0)}\,)W^{(1)}\,)W^{(2)}$ contains the embedding expression of semantic information fusion of documents, words and topics without direct connection. In the heterogeneous graph with three types of nodes, three-layer GCN can transfer information between nodes up to three steps away, so as to realize the information exchange between documents and topics without direct edges in the graph. In the preliminary small sample experiment, the two-layer GCN makes no information transmission between some documents and topics and their similar unconnected nodes, while the three-layer GCN will perform three propagation steps in the forward transmission process and

effectively convolute the third-order neighborhood of each node to realize information exchange.

IV. EXPERIMENT

A. Dataset

In order to verify the effectiveness of the model, this paper selects three data sets for experimental verification, namely, two public benchmark English text data sets 20NewsGroup and AGNews, and a self-collected Chinese dataset Ch_News. Among them, the English data set 20NewsGroup collected about 18,846 newsgroup documents, which were divided into 20 newsgroups sets with different topics, including three kinds of data information-document content, document title and document attribution tag. And there were no duplicate documents in the corpus. The English data set AGNews was published by Cornell University in 2004. It contains nearly one million news articles, which are mainly divided into four categories-World, Sports, Business and Sci/Tec. 8000 news articles were selected as corpus set in this paper. And the data volume of the four categories of news in these 8000 news articles is equal. The Chinese dataset Ch_News is made by manually complied and tagged about 1,000 news texts from Ecns.cn and english.chinamil.com.cn. The data come from five categories, namely, culture and education, finance and economics, military equipment, biotechnology and sports.

For Chinese and English datasets, different processing is carried out according to their language characteristics. For two English datasets, there are spaces between words to segment, and the preprocessing of removing low word frequency and stop words is directly carried out. For the Chinese dataset, according to the characteristics that there is no space to separate two characters in Chinese and the news contains a large number of domain knowledge entities, an entity dictionary is built, which contains the knowledge nouns in each domain in the dataset, and then the preprocessing of removing stop words, Chinese word segmentation and low-frequency words is carried out in turn.

Table 1 summarizes the statistics for the three data sets. Each data set takes 70% of the data as the training set, 30% as the test set, and then randomly selects 10% of the data in the training set as the verification set.

B. Comparative experiment

In order to verify the effect of the model on the text classification task, experiments are carried out based on tensorflow and keras, and the programming language is python3.6. In this paper, SVM+LDA, TextCNN and Text-GCN are used to compare with the proposed method. In the experiments of the three benchmark models, the parameters set in the original paper or the default parameters are used to reproduce. In the experiments of the three benchmark models, the parameters set in the original paper or the default parameters are used to reproduce. The three algorithms are briefly introduced as follows.

TABLE I. STATISTICAL TABLE OF DATA SETS USED IN EXPERIMENTS

Dataset	Total_docs	Training_docs	Testing_docs	Total_words	Categories
20NewsGroup	18,846	13,192	5,654	42,739	20
AGNews	8,000	5,600	2,400	32,653	4
CH_News	1,080	756	324	13,324	5

SVM + LDA The method extracts features through LDA topic model and uses SVM classifier to realize text classification [13].

TextCNN The method uses convolution neural network[21], using randomly initialized word embedding vectors as input, convoluting text matrix with filters of different lengths, and then using max pooling to extract vectors for each filter.

Text-GCN[8] The model construct word co-occurrence heterogeneous graph as graph model, and use graph convolution neural network to jointly learn the embedding of words and documents. After convolution, the feature representation vector learned for each node is obtained, which implies the prediction label of each document, so as to realize classification.

C. Result comparison and analysis

Text classification is a basic and classical task with classical evaluation parameters. Considering that this model solves the problem of multiple text classification. Accuracy is selected as the evaluation index in this paper, and its calculation method is as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

Where TP represents the number of correctly identified positive samples, FP represents the number of incorrectly identified positive samples, FN represents the number of incorrectly identified negative samples, and TN represents the number of correctly identified wrong samples.

Table 2 shows the accuracy test of text classification task for three data sets using different methods. By comparison, the method proposed in this paper has achieved better classification results on three data sets, especially in the self-constructed Chinese data set, which proves the effectiveness of the model.

TABLE II. COMPARISON OF EXPERIMENTAL RESULTS

DataSets / Model	20newsgroup	AGNews	Ch_News
SVM+LDA	0.6827	0.7241	0.6956
CNN	0.7160	0.8059	0.6893
Text-GCN	0.8634	0.9245	0.7402
LDA-THG+GCN	0.8796	0.9310	0.7594

The main reasons why using LDA-THG and graph convolution method is superior to the other three methods are: 1) text graph can capture the self-information of document and subject related auxiliary information at the same time; 2) GCN can well capture the information between high-order neighborhood nodes, and can calculate the new characteristics of nodes as the weighted average of themselves and their third-order neighbors. And the document and topic nodes in LDA-THG integrate the potential information of each other and enrich the expression of text features.

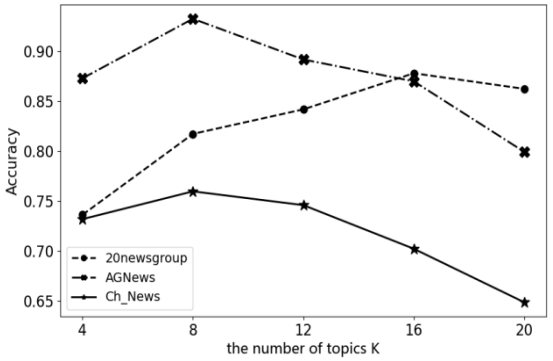


Figure 4. Influence of topic number on classification results of LDA-THG+GCN model on three data sets

The impact of different topic numbers K on the classification results in three data sets is shown in Figure 4. It can be seen that for the three datasets, the categories of the datasets themself are 20, 4

and 5, respectively. The range of their high classification accuracy is when K is [14,20], [4,12] and [4,12], that is, when K value is within the range corresponding to the category of the data set itself, the accuracy is relatively high. Therefore, when the number of topics is maintained in a specific range, the classification accuracy is high. When the number of topics K exceeds this range, it will decrease in varying degrees. The main reason for this phenomenon is that when there are too many or too few topics, the found topics will be very rough or too detailed, both of which will cause the document to lose some relevant topic information. At the same time, the grammatical and expressive features of the text of English and Chinese lead to a great difference in the classification accuracy of AGNews and Ch_News under similar circumstances.

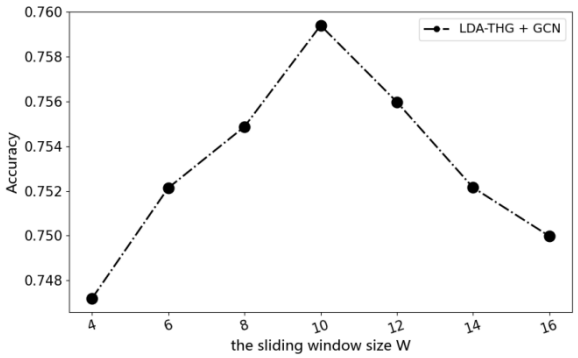


Figure 5. Influence of W on classification results of LDA-THG+GCN model on Ch_News data set

On Chinese data set Ch_News, the influence of the sliding window size W of this model on classification results is shown in Figure 5. As can be seen from the figure, with the increase of the sliding window, the accuracy rate will be improved first and then slowly decrease. Accuracy is highest when the window length is equal to 10. This shows that when the sliding window is too small, sufficient global word co-occurrence information can not be generated to represent the probability relationship of word co-occurrence, while when the sliding window is too large, edge relations will be added between the nodes that are not close, making the information to be trained redundant. This not only increases the difficulty of training and storage, but also makes the features in the text representation fail to highlight the main features contained in the text itself.

V. CONCLUSION

In this paper, a new method of news text classification based on graph convolution neural network is proposed by combining text topics and words in text. This method uses the topic distribution and the word distribution corresponding to the topic in the LDA model to construct the weights of nodes and edges in the graph model, and generates a heterogeneous network information graph. Then, the point and edge information of the graph is integrated through the multilayer graph convolution neural network and *sofmax* layer, and finally the label of the prediction document is obtained. The purpose of this method is to provide more reference knowledge for improving the performance of the model in text classification by integrating the data that helps to understand the text. Future research may further improve the performance of the model and achieve better training results by expanding the number of node layers involved in primary graph convolution, such as deep reasoning of two-layer nodes.

REFERENCES

[1] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, “Deep learning--based text classification: a comprehensive review,” ACM Computing Surveys (CSUR), vol. 54(3), April. 2022, pp. 1-40, doi:10.1145/3439726.

[2] H. Cai, VW Zheng, KCC Chang, “A comprehensive survey of graph embedding: Problems, techniques, and applications,” IEEE Transactions on Knowledge and Data Engineering, vol. 30(9), Sept. 2018, pp 1616-1637. doi:10.1109/TKDE. 2018.2807452.

[3] PW Battaglia, JB Hamrick, V. Bapst, et al. “Relational inductive biases, deep learning, and graph networks,” arXiv preprint, 2018. doi:10.48550/arXiv.1806.01261.

[4] D. Shen, G. Wang, W. Wang, et al. “Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms,” arXiv preprint, 2018, doi:10.48550/arXiv.1805.09843.

[5] Chen Y, “Convolutional neural network for sentence classification,” University of Waterloo, August 2015, http://hdl.handle.net/10012/9592.

[6] S. Lai, L. Xu, K Liu, J. Zhao, “Recurrent convolutional neural networks for text classification,” Twenty-ninth AAAI conference on artificial intelligence, 2015.

[7] R. Mihalcea, P. Tarau, “Textrank: Bringing order into text”， Proceedings of the 2004 conference on empirical methods in natural language processing，2004, pp 404-411.

[8] L. Yao, C. Mao, Y. Luo, “Graph convolutional networks for text classification,” Proceedings of the AAAI conference on artificial intelligence, vol. 33(01), July 2019, pp 7370-7377, doi:10.1609/aaai.v33i01.33017370.

[9] Y. Zhang, X. Yu, Z. Cui, et al. “Every document owns its structure: Inductive text classification via graph neural networks,” arXiv preprint, April 2020, doi:10.48550/arXiv.2004.13826.

[10] X. Liu, X. You, X. Zhang, J. Wu, P. Lv, “Tensor graph convolutional networks for text classification,” Proceedings of the AAAI conference on artificial intelligence, vol. 34(05), April 2020, pp 8409-8416, doi:10.1609/aaai.v34i05.6359.

[11] TN Kipf, M. Welling, “Semi-supervised classification with graph convolutional networks,” arXiv preprint, September 2016, doi:10.48550/arXiv.1609.02907.

[12] X. Zhang, J. Zhao, Y. LeCun, “Character-level convolutional networks for text classification,” Advances in neural information processing systems, 2015.

[13] DM Blei, AY Ng, MI Jordan, “Latent dirichlet allocation,”. Journal of machine Learning research, January 2003, pp 993-1022.

[14] R. Wang, Z. Li, J. Cao, T. Chen, L. Wang, “Convolutional recurrent neural networks for text classification,” 2019 International Joint Conference on Neural Networks (IJCNN), IEEE , September 2019, pp 1-6,doi:10.1109/IJCNN.2019.8852406.

[15] Y. KIM, “Convolutional neural networks for sentence classification,” 2014 Conference on Empirical Methods in Natural Language Processing, 2014, pp 1746-1751.

[16] Q. Li, H. Peng, J. Li, et al. “A survey on text classification: From shallow to deep learning,” arXiv preprint, August 2020, doi:10.48550/arXiv.2008.00364.

[17] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, D. Brown, “Text classification algorithms: A survey,”. Information, vol 10(4), October 2019, pp 150, doi:10.3390/info10040150.

[18] M. Malekzadeh, P. Hajibabaee, M. Heidari, S. Zad, O. Uzuner, J. H Jones, “Review of graph neural network in text classification,” 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE, December 2021, pp 0084-0091, doi:10.1109/UEMCON53757.2021.9666633.

[19] S. Ghosh, S. Maji, MS. Desarkar, “Graph Neural Network Enhanced Language Models for Efficient Multilingual Text Classification,” arXiv preprint, March 2022, doi:10.48550/arXiv.2203.02912.