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: December 28, 2021

Editor in Chief

Associate Editor-in-Chief

Administrator

Associate Editors

Language Editor

Table of Contents

Research on the Gaze Direction of Head-Eye Data Fusion

Xin Xu*
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710021, Shaanxi, China
E-mail: Ariel970504@163.com
*corresponding author

Changyuan Wang
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, 710021, Shaanxi, China
E-mail: Cyw901@163.com

***Abstract*—The line of sight refers to the gaze direction of the human eye and reflects the focus of human attention. Head movement is an important accompanying behavior in the process of human gaze, and it is of great significance to human visual attention. This paper intends to combine the gaze focus calculation modeling of head movement and eye movement combined with deep learning data fusion. By combining the gaze direction calculation model of data fusion and neural network algorithm, deep learning technology is used to reveal the relationship between head movement and eye movement, and the data of head movement and eye movement are merged to realize accurate and fast real-time gaze spatial direction calculation. New ideas for improving the efficiency, reliability, usability and functionality of the gaze tracking system. In this paper, the convolutional neural network method is used, and the classification accuracy of the line of sight direction reaches 99% when the head posture is free.**

Keywords-Head Pose Estimation; Pupil Center Detection; Line of Sight Direction; Data Fusion; Neural Network

I. INTRODUCTION

The cockpits of modern advanced fighter jets mostly adopt a one-level three-down display mode. With the development of avionics technology, the down-view display gradually adopts a large-size overall touch screen, which improves ergonomics by integrating visual information. The large number of manual operations that still exist are an important factor that limits the current efficiency of human-computer interaction. The gaze tracking technology obtains the direction of the gaze by measuring the positioning and posture of the human eye. Based on human physiological characteristics, the gaze response is fast and accurate, and is not affected by high overload. It has obvious advantages in the field of aviation human-computer interaction.

A. Eye tracking technology

Eye tracking technology refers to the use of certain features that are relatively unchanged during eye movement to obtain pupil data, including pupil center coordinates, contours and other parameters. Due to the superior performance of deep learning technology in the field of computer vision, currently, eye tracking technology is mainly developed around this technology. Fuhl et al. [1] proposed a coarse-to-fine pupil detection model-pupilnet based on convolutional neural networks. The team input the eye pictures into the rough recognition network in blocks, and the area with the highest score is the pupil. Rough area, and then input it into the precise recognition network to extract pupil parameters. Due to the strong dependence of the neural network on the data set, the unbalanced distribution of the image types in the data set will cause the model to have larger detection errors for the relatively small image types . Based on this feature, Shaharam Eivazi et al. [2] proposed a targeted image enhancement method for pupil detection errors caused by mirror reflection, sunlight reflection, blurring, etc., and used the enhanced image as a data set to train classic convolution The neural network recognition model achieves the best simultaneous detection model. Fuhl et al. [3] proposed to use Cycle GANs [4] to enhance the image, increase the richness of the data set, and perform pupil segmentation on it. The segmentation result only retains the shape and position of the pupil, and filters out all the noise,

DOI: 10.21307/ijanmc-2021-031

thereby improving the pupil Detection accuracy. The above models are all based on the pupil characteristics of a single frame image, which is also a characteristic of convolutional neural networks. It fails to introduce the pupil movement characteristics embodied in consecutive frames. At the same time, the high complexity of the model makes it difficult to apply to real-time tracking tasks. In order to solve the above Two problems, this project intends to introduce Long Short-Term Memory (LSTM) [5], use its ability to process continuous sequences to introduce pupil movement features to improve recognition accuracy, and add a "pruning" operation [6] Reduce model calculation speed to achieve real-time tracking performance.

B. Head pose estimation

Head pose refers to the use of computer vision and pattern recognition technology to estimate the orientation of the head in a digital image. The head movement data mainly includes the three-axis spatial position and the three-axis posture. Among them, the three-axis posture data is composed of three sets of yaw, pitch and roll data, reflecting the head space rotation state. Image-based head motion tracking mainly uses facial features as reference points to match the three-dimensional model of the face through the recognition of the head and face.

Bao et al. [7] proposed using a three-layer convolutional neural network for head pose estimation, and used a coarse-to-fine method in the model training process. For the application of deep learning in the field of head pose estimation, Patacchiola et al.[8] conducted experiments on four convolutional neural networks and showed that the shallow architecture shows better performance on small-scale data sets, while the deep architecture is more Suitable for large-scale data.

C. Line of sight estimation

The line of sight estimation mainly solves the nonlinear mapping relationship between the collected human body information and the line of sight. Based on the difference of the collection equipment, it is mainly divided into wearable and non-wearable line of sight estimation methods. For wearable devices, Thiago Santini et al. [9] used glasses-type acquisition devices and assumed that the gaze tracker was still a rigid body after calibration. On this basis, the tracker coordinates were used to quickly estimate the gaze angle, and then a second-order polynomial regression model was used. Mapping it to the two-dimensional field of view image collected by the tracker to obtain a precise landing point. Although the model has a good detection effect on its experimental data set, its dependence on the pupil contour leads to occlusion and other interference which will cause a larger line of sight Estimated error.

For non-wearable devices, if the freedom of the user's head movement is restricted, the line of sight direction can be estimated based on facial features alone. Su Haiming et al. [10] proposed to determine the position of the human eye through a face positioning algorithm and use the pupil template to detect it. The pupil area is then used to establish the gaze point mapping relationship using the neural network model. When the head posture is fixed, the recognition accuracy rate of 96.74% is achieved. This type of method limits the range of head movement. Although it has achieved high recognition performance, it can only be applied to specific fields. Regarding the line of sight estimation in free pose, S Park et al. [11] pointed out that in this case, high-precision line of sight estimation based on the eye picture is not suitable for the task. The spatial position of the pupil center of the human eye that determines the direction of the line of sight is not observable in the picture. . Therefore, this type of method adds head posture parameters to the eye movement image technology to establish a head-eye-line-of-sight model. Xucong Zhang et al. [12] proposed GazeNet, which takes normalized eye pictures and head pose parameters as parameters, uses VGG16 [13] as the basic network model to extract the feature spectrum of the eye pictures, and cascades the head pose angle to The output vector of the first fully connected layer of the line-of-sight estimation network is compensated, and the line-of-sight angle is estimated through the line-of-sight estimation network.

Based on the existing research foundation, it is found that in the human visual gaze system, head movement expands the person’s field of vision. At the same time, if the head space movement and posture cannot be clarified, the eye movement coordinate system cannot be clarified. Therefore, this article starts from the non-wearable head movement and eye movement detection and tracking, through image recognition measurement, head space movement and eye movement measurement, combined with data fusion and deep learning neural network algorithm technology to establish a line of sight calculation model, Clarify the mathematical relationship between head movement, eye movement and the direction of the line of sight, and realize the acquisition of the direction of the binocular line of sight.

II. RELATED WORK

A. Based on non-wearable multi-lens camera data acquisition and simulation flight experimental platform construction

In this paper, a non-wearable device is used to collect data from the head and eyeballs, that is, to set up multiple cameras and infrared light sources in the environment to collect images of the head, face and eyeballs. The control system is divided into upper computer and lower computer. The upper computer controls the multi-eye camera and infrared light source through the control chip. The lower computer is responsible for processing the image data and measuring the direction of sight. At the same time, the simulated flight platform is improved. Based on the original six-axis full-motion platform, according to the design of the J10 cockpit, a head-up display, a down-view display and a visual simulation display are added. The simulated flight experiment platform is shown in Figure 1.

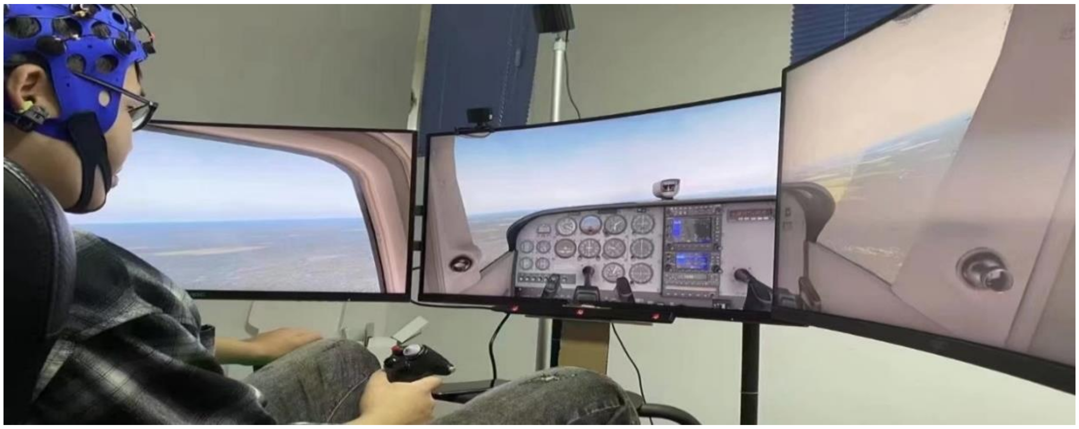


Figure 1. Flight simulation platform

B. Head movement measurement

Head movement measurement refers to measuring the movement trajectory of the subject's head relative to an absolute reference point (usually the spatial coordinates of the back of the head, forehead center or neck in the initial sitting position) in real or virtual space, and calculating the movement paradigm to obtain The mathematical model of the environment where the research object is located [14]. In the actual environment, it is necessary to pay attention to the vector components of the three-dimensional orthogonal coordinate system with the calibration point as the origin, and use the orthogonal vector to describe the movement characteristics of the head; and for the head movement measurement in the virtual space, the two-dimensional image is mainly used. , One of the vectors needs to be converted into a head depth function to obtain the motion behavior in the actual environment through the two-dimensional image, as shown in 2.

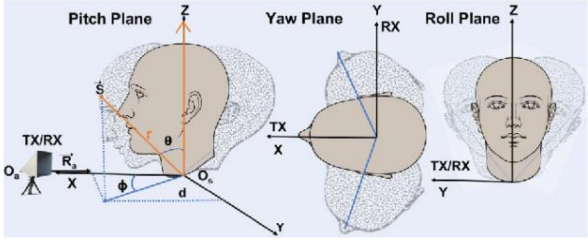


Figure 2. Spatial coordinates of head movement measurement (neck origin)

With the rapid development of computer vision and the decline in the price of motion sensor hardware, many image and sensor-based head motion measurement solutions have also been proposed. As a part of the human torso, the development of motion measurement and the field of motion capture of the head are advancing almost at the same time [15-17], and head motion measurement has driven the development of human-computer interaction, assisted driving, and machine-human collaboration; similarly, Since the head image contains information such as eyes, facial expressions, and facial features, head movement measurement is also used in live body detection, focus tracking, and target recognition. At present, the more accurate solution in the above-mentioned applications is motion measurement based on motion sensors, the more common is motion measurement based on image displacement algorithms, and the more cutting-edge is image-based deep learning motion measurement.

The sensor-based motion detection scheme is based on the motion sensor placed on the head. The head motion is obtained by calculating the relative position of the current head's spatial position and the initial position (or calibration point) through the acceleration and angular velocity vectors built into the sensor. Way. Existing hardware can easily reach the sampling rate of 30Hz and above. The movement track can be obtained by recording the displacement of continuous time. Taking the sensor hardware used in this article as an example, its appearance is shown in 3.

Figure 3. Xsens® MTi-G motion sensor used in the article

The known acceleration is shown in Eq. (1). The natural coordinate system decomposes to obtain the tangential acceleration and the normal acceleration, and the acceleration decomposes in the circular motion to obtain the tangential acceleration and the centripetal acceleration.

$$\alpha = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = \frac{dv}{dt} \tag{1}$$

When studying the problem, the sensor fixed on the head can be regarded as a mass point in space, then the sensor acceleration is decomposed into a_t tangential acceleration and a_n normal acceleration, as shown in Eq. (2).

$$a = a_t + a_n \tag{2}$$

Furthermore, the acceleration modulus, angular velocity vector, and angular acceleration vector shown in Eq. (3) to Eq. (5) are obtained.

$$|a| = \sqrt{a_t^2 + a_n^2} \tag{3}$$

$$\omega = \lim_{\Delta t \to 0} \frac{\Delta n}{\Delta t} = \frac{dn}{dt} \tag{4}$$

$$\alpha = \lim_{\Delta t \to 0} \frac{\Delta \omega}{\Delta t} = \frac{d\omega}{dt} \tag{5}$$

According to the vector obtained above, the trajectory of the mass point (the center of gravity of the sensor) in space can be obtained as shown in Figure 4. Figure 5 shows the research process of head movement measurement based on motion sensors.

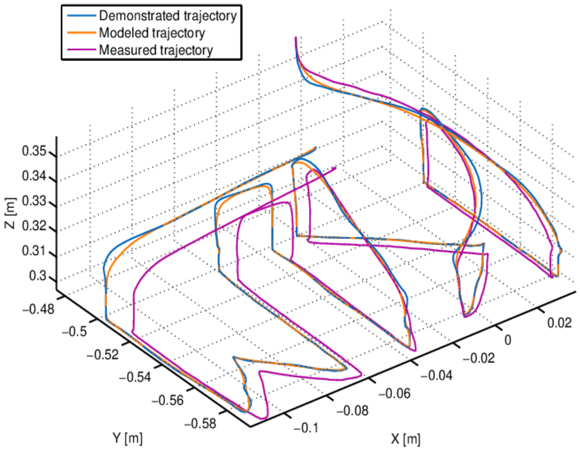


Figure 4. Schematic diagram of particle motion in space

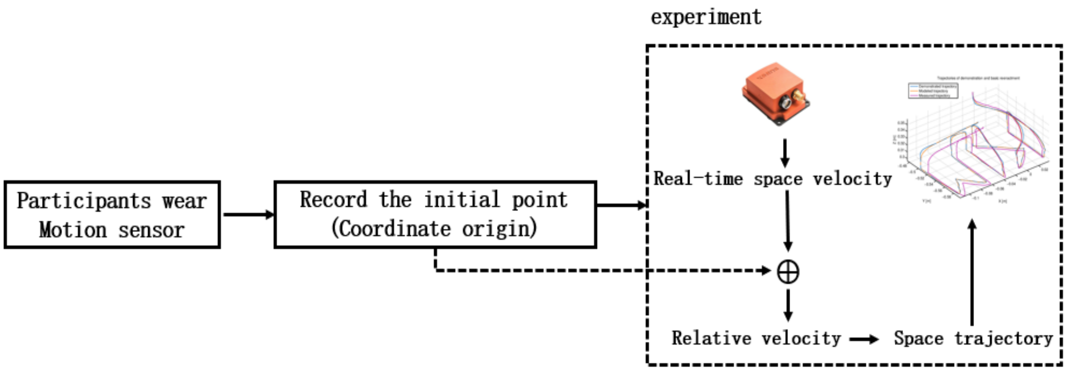

experiment
Participants wear Motion sensor
Record the initial point (Coordinate origin)
Real-time space velocity
⊕
Relative velocity
Space trajectory

Figure 5. Experiment process of head movement measurement based on motion sensor

C. Eye movement detection

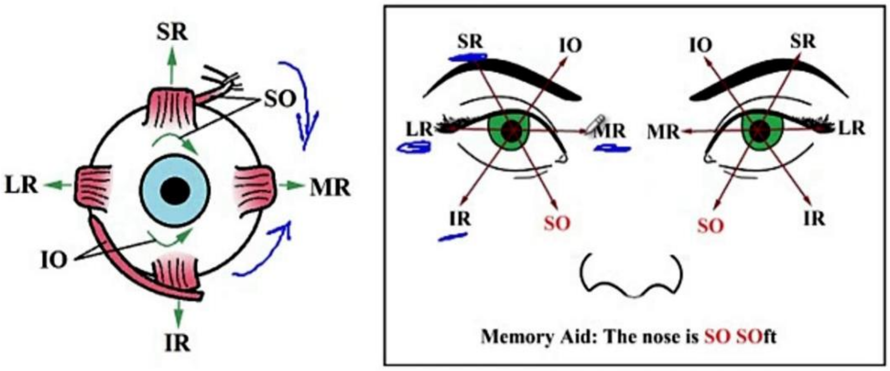


Figure 6. Eye movement measurement plane coordinates (origin of nose tip)

Eye movement measurement refers to measuring the relative movement of the eyeball and this point with a relative reference point as the origin of the coordinates (usually the center of the brow and the tip of the nose). Because the eyeball is restricted to close to two degrees of freedom by the human skull structure, four straight eyeball muscles and two oblique muscles [18], the displacement of the eyeball perpendicular to the longitudinal section of the skull is less than 10 % [19-21]. At the same time, most of the tasks of eye movement are to quantify the center of the pupil and the point of sight. The measurement of the task can be simplified into a two-dimensional domain, as shown in 6.

Eye movement measurement is different from head movement measurement, which requires the algorithm to have higher calculation accuracy. There are two main types of eye movement measurement methods: measurement based on head-mounted infrared camera and measurement based on non-contact infrared camera. The reason why the infrared camera is selected as the eye movement capture device is that the reflection effect of the human eye pupils on infrared wavelengths is better than that of visible light [22-24]. Figure 7 shows the study of pupil reflectivity at different infrared wavelengths by scholars; Figure 8 compares the imaging effects of RGB cameras and infrared cameras on eye features. In the measurement range of the eyeball diameter, the state of pupil zoom cannot be easily ignored, especially in terms of gaze tracking, the pupil zoom represents the change of the focus of the line of sight, but also the change of the range of attention [25].

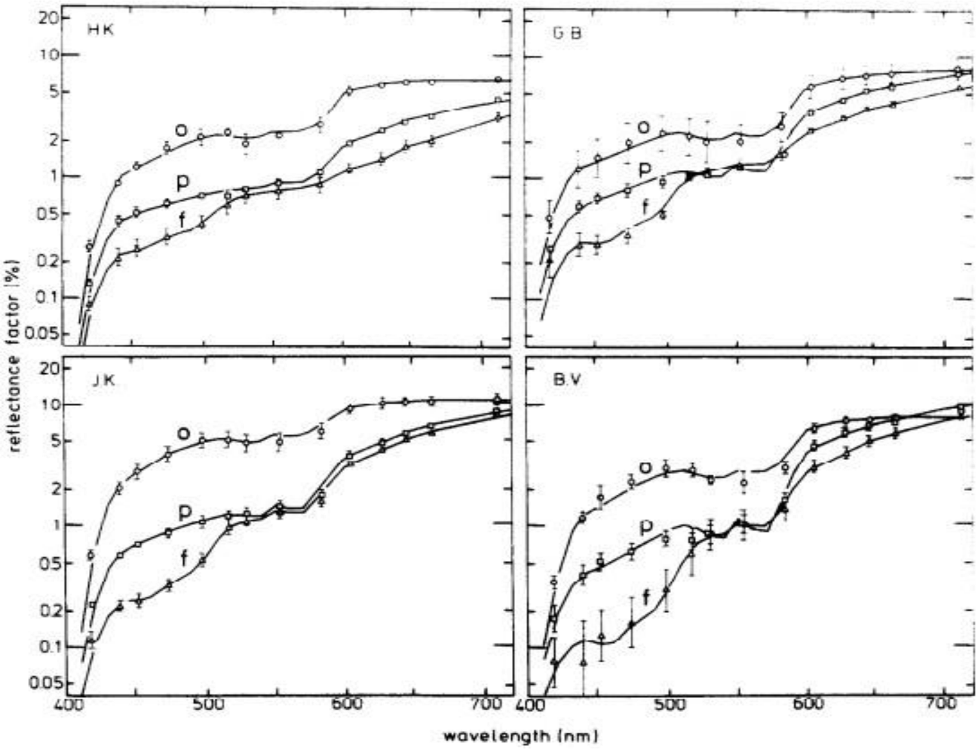


Figure 7. Pupil reflectance curve at different wavelengths

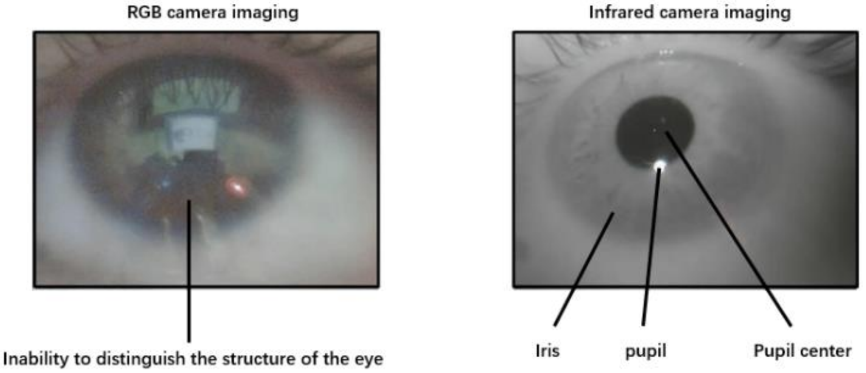


Figure 8. Comparison of pupil imaging between RGB camera and infrared camera

Because the head space position and posture data can clarify the eye movement coordinate system, they are indispensable data in the calculation of the line of sight direction. This paper intends to fuse head space movement and eye movement data, establish a gaze direction calculation model combining data fusion and deep learning neural network algorithm, and conduct machine learning training neural network through calibration experiments to clarify the relationship between head movement and gaze direction. Realize the measurement of the line of sight.

III. TECHNICAL ROUTE

This article puts forward requirements for multiple indicators such as head movement eye movement measurement accuracy, field of view, device portability, non-contact conditions, and rapid output of results. There is currently no such highly integrated research reference and hardware conditions in related fields , Especially the field of view in the head movement measurement and the calculation accuracy in the eye movement measurement. These two indicators are the technical difficulties and innovative entry points of this article. The portability of the equipment requires that the entire system can be easily and quickly deployed on various platforms, that is, to reduce the proportion of customized equipment; non-contact conditions require the use of the testees to reduce the burden and learning costs; the rapid output of the results requires that the text can be implemented, the traditional The image processing method takes a lot of time to calculate the head movement, and involves the solution of a large number of nonlinear equations.

The motion sensor can accurately capture the motion trajectory, speed, acceleration and torsion angle of the participant's limbs in space. These data will provide accurate labels during the text construction stage to modify the performance of the model. The trinocular infrared camera layout is selected as the only hardware in the application stage of this article. In the application process, the motion sensor will no longer be used, but the infrared images of different angles collected by the trinocular camera will be directly calculated.

Deep neural networks have developed rapidly in recent years. One of the characteristics is that training models are time-consuming and the amount of calculation in application is generally less than traditional methods. This is because the nature of deep neural networks is to fit piecewise nonlinear functions through multiple linear equations. The concept of gradient, divergence, etc. is converted to linear operations. This attribute determines that deep neural networks will not generate a large number of partial differential equations at the application stage, but instead are alternative multiplication operations. This article chooses deep neural network as a tool, responsible for deriving the technical difficulties mentioned above.

In summary, the technical route of the construction phase of this article is shown in Figure 9, and the process of embedding the motion paradigm into the deep neural network model is shown in Figure 10.

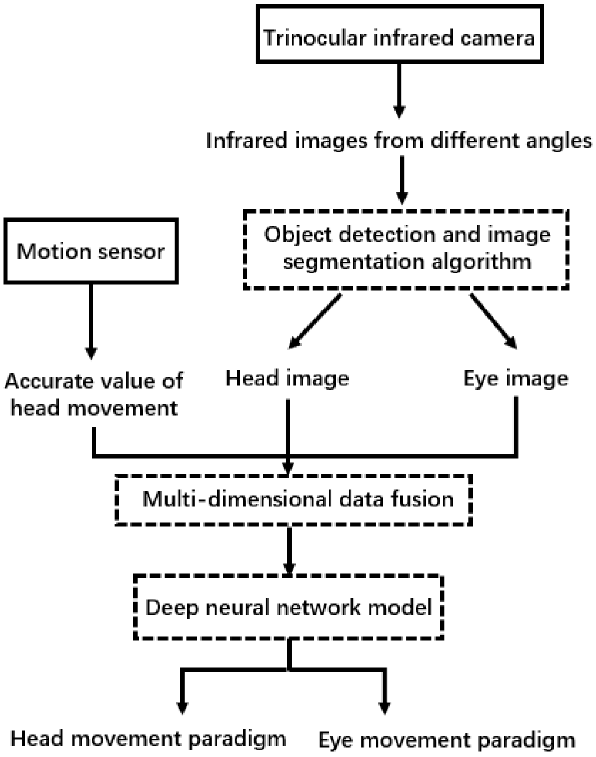

Trinocular infrared camera
Infrared images from different angles
Motion sensor
Object detection and image segmentation algorithm
Accurate value of head movement
Head image
Eye image
Multi-dimensional data fusion
Deep neural network model
Head movement paradigm
Eye movement paradigm

Figure 9. Technical route

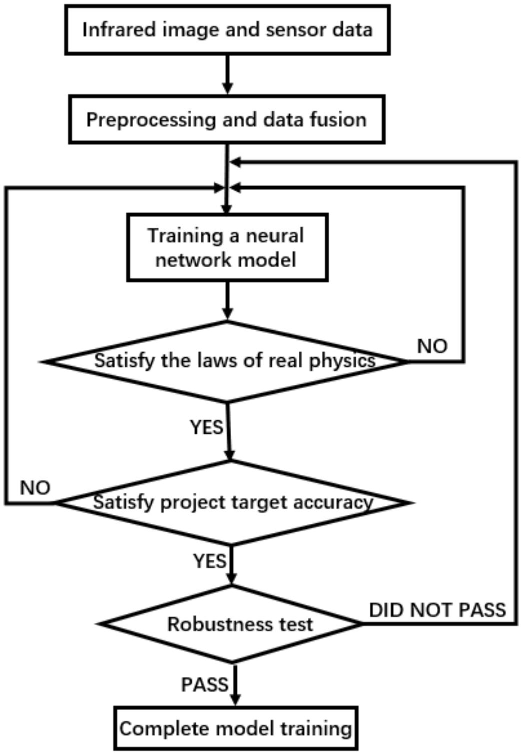

Infrared image and sensor data
Preprocessing and data fusion
Training a neural network model
Satisfy the laws of real physics
NO
YES
NO
Satisfy project target accuracy
YES
Robustness test
DID NOT PASS
PASS
Complete model training

Figure 10. Model training process

IV. EXPRIMENT

A. Data collection

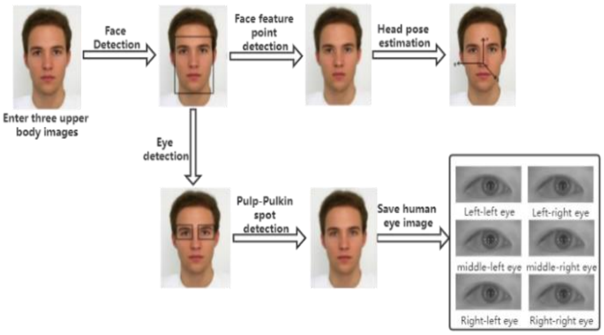


Figure 11. Training set collection process

The data set in this paper is collected in the laboratory, and the process of collecting eye images and head posture is shown in Figure 11.

The data collection process in this article is as follows: A real-time data collection program is implemented, three pure black pictures of 2560×1440 size are displayed on the full screen, and a red circle appears at a random position on the screen, and the image collector is required to look at the red circle target and tap Space bar, this is to ensure that the collector is focused. There are no other requirements for the image collector, and the head can move freely. While gazing at the point, press the space bar, the program will save the three upper body photos of the image collector at that moment, and save the corresponding head posture and gaze point to a text file (one-to-one correspondence with the number), and then click Another red circled target appears at a random location on the screen. The fixation point is the coordinate value of the red circle target randomly appearing on the computer. The 3D coordinates of the fixation point can be determined by the camera posture and the position of the computer screen in the world coordinate system. The direction of the line of sight is the distance from the 3D coordinates of the eye to the 3D position of the fixation point Connect. A schematic diagram of the data collection process is shown in Figure 12, and part of the collected face pictures are shown in Figure 13.

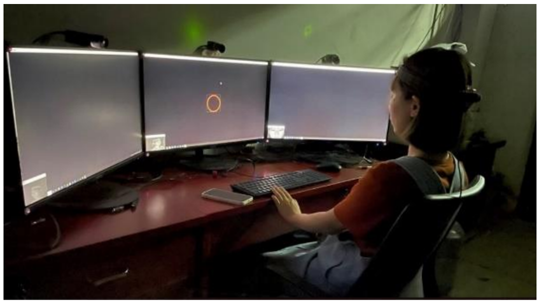


Figure 12. Data collection diagram

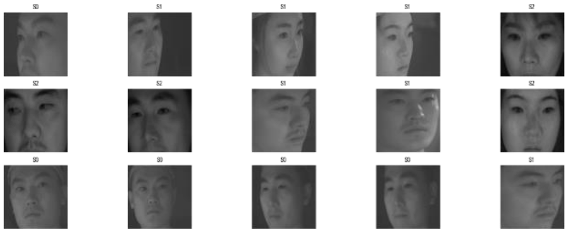


Figure 13. Part of the data set display

B. Facial feature recognition

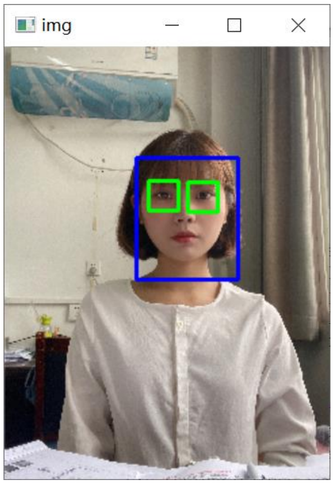


Figure 14. Facial feature recognition

After comparative research, the Adaboost algorithm is selected for face detection. The Adaboost algorithm selects the best features from a large number of Haar features and converts them into weak classifiers for classification and use, so as to achieve the purpose of classifying the target. The face recognition result based on the Adaboost algorithm is shown in Figure 14.

C. Human eye feature recognition

The facial features have been successfully recognized through the above process, and the next step is to recognize the eye area. Recognize the eye area based on the Adaboost algorithm. As shown in the figure, the white point indicated by 15 is the light spot (Pulchin spot) formed on the human cornea by the near-infrared light source on the screen, and the darkest area in the center of the human eye area pointed to by 2 is the human eye pupil area. The slightly shallower area pointed to by 3 is the iris area of the human eye.

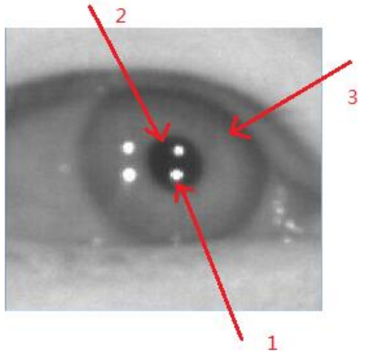


Figure 15. Eye area feature map

Take a screenshot of the identified eye area. The three photos at the same time are divided into one group, each group gets a total of 6 eye photos, and the pupil-Pulchin spot picture can be clearly seen, as shown in Figure 16.

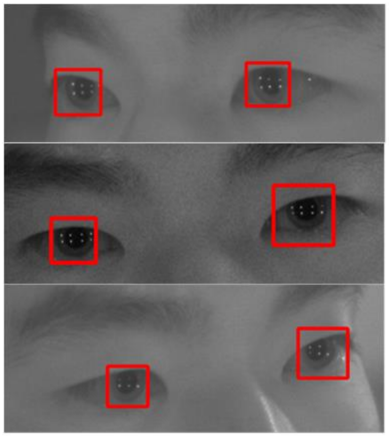


Figure 16. Accuracy comparison chart

D. *Training and testing*

The two human eye images in the data set collected by this article correspond to a head posture and line of sight direction (assuming that the left eye and the right eye have the same line of sight direction).

The convolutional neural network model used in this article is trained under the deep learning open source framework caffe. The training model requires two configuration files: solver. prototxt and train test. Prototxt

The solver. prototxt mainly includes the setting of training parameters, including the number of iterations, weight attenuation coefficient learning rate, impulse, display test error iteration interval, GPU and CPU settings, etc. The parameter selection in this paper selects the optimal parameters through 10-fold cross-validation, and then retrains all the training sets with the selected parameters to obtain the line-of-sight estimation model.

Train_test. prototxt mainly includes the settings of the convolutional neural network structure, specifically the directory of training data and test data, convolutional layer and pooling layer settings, fully connected layer settings, loss function, pooling type, etc.

On the data set collected by myself, the accuracy graphs of the training set and the test set are shown in Figure 17, where the horizontal axis represents the number of iterations, the blue line represents the training error, and the orange line represents the test error.

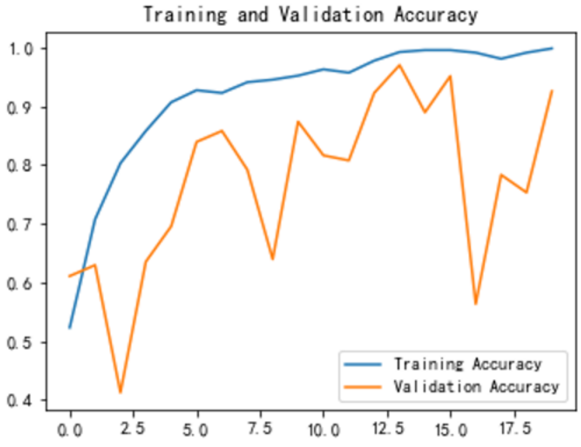


Figure 17. Pupil-Pulchin spot image recognition

The iterative loss curve of training on the self-collected data set is shown in Figure 18, where the horizontal axis represents the number of iterations, the blue line represents the training error, and the orange line represents the test error.

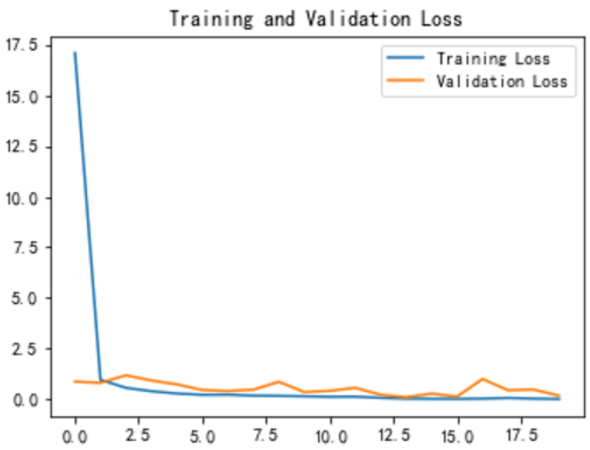


Figure 18. Loss curve

Confusion matrix, also known as error matrix, is a standard format for precision evaluation. Through the confusion matrix, we can clearly see the number of correct identifications and the number of incorrect identifications in each category in the three directions, and then quickly help us analyze the misclassification of each category.In this paper, the convolutional neural network method is used. When the head posture is free, the classification accuracy of the line of sight direction reaches 99%. The classification result is shown in Figure 19.

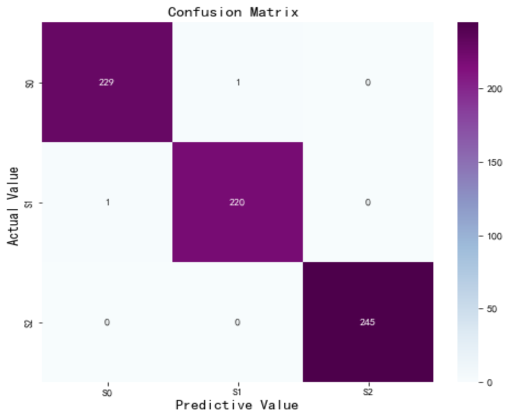


Figure 19. Confusion matrix

V. CONCLUSION

Sight tracking is widely used in psychology, graphics, software engineering, pattern recognition, human-computer interaction, medicine, advertising psychology, military and many other fields. It has strong practical value. Therefore, gaze tracking has become a computer vision and pattern in recent years. Hot topics in the field of identification.In this paper, the pupil-Pulchin spot image method is used to measure the eye movement by the non-wearable image measurement method, combined with the head posture data, through the data fusion combined with the neural network algorithm of the line of sight calculation model, and the deep learning technology is used to reveal the head movement. The relationship with the gaze direction, fusion of head movement and eye movement data, to achieve accurate and fast real-time gaze spatial direction calculation. In this paper, the convolutional neural network method is used, and the classification accuracy of the line of sight direction reaches 99% when the head posture is free.

REFERENCES

[1] Fuhl W, Santini T, Kasneci G. Pupilnet: Convolutional neural networks for robust pupil detection[J]. arXiv preprint arXiv:1601.04902, 2016.

[2] Eivazi S, Santini T, Keshavarzi A. Improving real-time CNN-based pupil detection through domain-specific data augmentation[C]. Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, 2019: 1-6.

[3] Fuhl W, Geisler D, Rosenstiel W. The applicability of Cycle GANs for pupil and eyelid segmentation, data generation and image refinement[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019: 0-0.

[4] Zhu J-Y, Park T, Isola P. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]. Proceedings of the IEEE international conference on computer vision, 2017: 2223-2232.

[5] Tsironi E, Barros P, Weber C. An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition[J]. Neurocomputing, 2017, 268: 76-86.

[6] Liu Z, Li J, Shen Z. Learning efficient convolutional networks through network slimming[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 2736-2744.

[7] Bao J, Ye M. Head pose estimation based on robust convolutional neural network[J]. Cybernetics and Information Technologies, 2016, 16(6): 133-145.

[8] Patacchiola M, Cangelosi A. Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods[J]. Pattern Recognition, 2017, 71: 132-143.

[9] Santini T, Niehorster D C, Kasneci E. Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking[C]. Proceedings of the 11th ACM symposium on eye tracking research & applications, 2019: 1-10.

[10] Su Haiming, Hou Zhenjie, Liang Jiuzhen. A gaze tracking method using geometric features of human eyes [J]. Journal of Image and Graphics, 2019(201906): 914-923.

[11] Park S, Spurr A, Hilliges O. Deep pictorial gaze estimation[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 721-738.

[12] Zhang X, Sugano Y, Fritz M. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 162-175.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv preprint arXiv:1409.1556, 2014.

[14] Mase K, Watanabe Y, Suenaga Y. Real-time head motion detection system. Sensing and Reconstruction of Three-Dimensional Objects and Scenes: International Society for Optics and Photonics; 1990. p. 262-8.

[15] Rolland JP, Davis LD, Baillot Y. A survey of tracking technologies for virtual environments. Fundamentals of wearable computers and augmented reality: CRC Press; 2001. p. 83-128.

[16] Zhou H, Hu HJBsp, control. Human motion tracking for rehabilitation—A survey. 2008;3:1-18.

[17] Al-Rahayfeh A, Faezipour MJIjoteih, medicine. Eye tracking and head movement detection: A state-of-art survey. 2013; 1:2100212.

[18] Von Lüdinghausen MJCATOJotAAoCA, Anatomists tBAoC. Bilateral supernumerary rectus muscles of the orbit. 1998; 11:271-7.

[19] Raudonis V, Simutis R, Narvydas G. Discrete eye tracking for medical applications. 2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies: IEEE; 2009. p. 1-6.

[20] Botha CP, de Graaf T, Schutte S, Root R, Wielopolski P, van der Helm FC, et al. MRI-based visualisation of orbital fat deformation during eye motion. Visualization in medicine and life sciences: Springer; 2008. p. 221-33.

[21] Glarin RK, Nguyen BN, Cleary JO, Kolbe SC, Ordidge RJ, Bui BV, et al. Mr-eye: high-resolution mri of the human eye and orbit at ultrahigh field (7t). 2021; 29:103-16.

[22] Uhl A, Wild P. Multi-stage visible wavelength and near infrared iris segmentation framework. International Conference Image Analysis and Recognition: Springer; 2012. p. 1-10.

[23] Loskutova E, Butler JS, Hernandez Martinez G, Flitcroft I, Loughman JJCER. Macular Pigment Optical Density Fluctuation as a Function of Pupillary Mydriasis: Methodological Considerations for Dual-Wavelength Autofluorescence. 2021; 46:532-8.

[24] Jan F, Usman IJO. Iris segmentation for visible wavelength and near infrared eye images. 2014; 125:4274-82.

[25] Chen Y, Davoine F. Simultaneous Tracking of Rigid Head Motion and Non-rigid Facial Animation by Analyzing Local Features Statistically. BMVC2006.

High Quality Network Connection and the Development of Internet of Things Drive the Demand of Wi-Fi-6

TANO
Daihatsu Motor Co., Ltd
E-mail: tano19210@gmail.com

Lu Yuyan
Photoelectrical Engineering Institute
Xi’an Technological University
Xi’an, China
E-mail: 592073729@qq.com

***Abstract*—As the most popular technology, Wi-Fi is built based on Wi-Fi standards in many consumer areas. Up to now, Wi-Fi has experienced more than 20 years of development. The launch of the first generation protocol IEEE802.11 has a low popularity because it is not competitive in transmission speed and distance. With the proposal of the third generation 802.1g protocol, IEEE adds the backward compatibility feature to the new protocol. For 802.11 protocol, new technologies will be adopted in each iteration, so as to improve the network performance. Until the latest generation protocol, IEEE802.11ax launched in September 2019 has been called Wi-Fi 6.**

Wi-Fi-6 has greatly improved its performance. It introduces uplink MU-MIMO, OFDM orthogonal frequency division multiple access, 1024-qam high-order modulation and other technologies to solve the problems of spectrum resource utilization and multi-user access. Even if Wi-Fi-6 continues every new WLAN standard for a long time and only makes the conventional upgrade of single stream throughput, the main motivation behind Wi-Fi-6 is to improve the experience quality of MU-MIMO users and achieve this goal by minimizing the overall delay. Therefore, higher spectrum and the effective and efficient sharing of the spectrum have become the key to achieve the minimum capacity.

This paper also puts forward some views on the cooperation between 5G and Wi-Fi-6. For Wi-Fi networks, it has a faster update cycle than mobile networks. In addition, when users' usage habits are formed, the possibility of Wi-Fi network being replaced is low. Although 5G and Wi-Fi-6 differ in performance, they have their own advantages. They are suitable for different scenarios. 5G will have better performance in outdoor open places, such as car networking, while indoor Wi-Fi network can meet the more economical demand for high-speed Internet access, such as VR, smart home, etc.

In the third chapter of this paper, it focuses on the demand of Wi-Fi-6 driven by the development of high-quality network connection and Internet of things.

First of all, when Wi-Fi-6 was designed, high-density and high-capacity network service support has been considered, which is also the reason for supporting Wi-Fi-6 to be used in outdoor large-scale public places, indoor high-density wireless office and electronic office. However, because 5g has more advantages in the application of outdoor environment, indoor is the key application scenario of Wi-Fi-6. It is also because people's demand for information consumption, entertainment and smart device links has promoted the development of Wi-Fi-6.

DOI: 10.21307/ijanmc-2021-032

Secondly, the demand for high-quality network connection is prominent. With the promotion of national policies, it will help to speed up the construction of HD video content. As the main carrier of indoor transmission, Wi-Fi will be popularized more effectively when the terminal penetration continues to improve.

Thirdly, the demand for the Internet of things is gradually increasing. Unlike users such as laptops, IOT devices need deterministic wireless services. Therefore, enterprise Wi-Fi-6 has increasingly become the preferred platform for the indoor Internet of things. Compared with Wi-Fi-5 and 4G, Wi-Fi-6 has faster speed and can transmit data like lightning. Moreover, Wi-Fi-6 has improved in security and scalability respectively. The previous Wi-Fi router supports 250 devices to connect at the same time, but this is insignificant in front of Wi-Fi-6 that can support up to 1024 devices at the same time. Therefore, Wi-Fi-6 is ideal and economical for families and small businesses.

Finally, Wi-Fi-6 realizes the intelligent industrial scene. Although 5g and Wi-Fi-6 can now increase the rate to about 10gpbs due to technological improvements. However, because 5g network construction has just started, Wi-Fi-6 currently has greater advantages in the industrial field. Most factories have generally adopted wired network connection to realize Wi-Fi communication.

In the fourth chapter, this paper also summarizes the rapid expansion of Wi-Fi-6 market.

Today, the WLAN market is in a steady upward trend, and the Wi-Fi-6 application scenario is getting better. It is expected that the Wi-Fi-6 market will reach 24 billion yuan in 2023, which means that the chips supporting Wi-Fi-6 standards account for nearly 90% of the total. In terms of market segments, the market scale of router / gateway Wi-Fi-6 chip in China is about 300 million yuan in 2019 and is expected to exceed 4.5 billion yuan in 2023.

In the past two years, the development speed of Wi-Fi-6 is obvious to all. From new mobile phones of major brands to routers, products supporting Wi-Fi-6 technology continue to emerge.

Keyword-WLAN; Wi-Fi-6; 5G; Future Network

I. WI-FI STANDARD CONTINUES TO EVOLVE

A. Wi-Fi is a technology to realize WLAN.

Wireless local area networks (WLAN) uses radio frequency technology and electromagnetic wave to replace network cable, so as to make up for the defect of wired network coverage and achieve the purpose of network extension.

Wi-Fi (wireless fidelity) is a communication technology that conforms to IEEE 802.11 series wireless network specifications and has mutual compatibility. By definition, Wi-Fi is a technology to realize WLAN. Although there are many technologies and standards to realize WLAN, Wi-Fi is the most mainstream technology at present because the common WLAN in the consumer field is built based on Wi-Fi standards.

B. Development history of Wi-Fi

Wi-Fi has been developed for more than 20 years. 802.11 protocol began in the 1990s. In the early 1990s, the Institute of electrical and Electronics Engineers (IEEE) established a relevant working group to study and formulate the standard keeping protocol of WLAN [1]. Over time, different versions of 802.11 protocol are iterated to meet the network requirements.

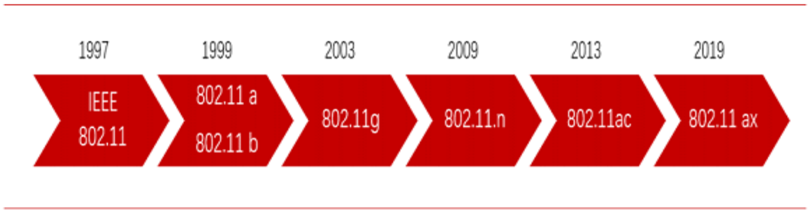


Figure 1. Important development history of 802.11 protocol

The first generation protocol IEEE 802.11-1997 was launched in June 1997, but its popularity is low because it is not competitive in transmission speed and distance. Later, IEEE 802.11 a / b protocol was introduced, in which 802.11 a protocol set the frequency band at 5GHz, and its maximum speed in the IOT layer was greatly improved to 54Mbps, but its development was limited due to slow chip development and other reasons; 802.11b protocol is based on 2.4GHz. Although its transmission speed is lower than 802.11a, it has better coverage and penetration. Because 802.11 a / b protocol was incompatible and B protocol was better in practicability, it occupied the market at that time.

802.11 g protocol is the third generation standard formulated by IEEE in July 2003. It integrates two versions of the previous generation protocol and can realize transmission at 2.4GHz and 5GHz. In addition, since this generation of protocols, IEEE has added the feature of backward compatibility to the formulation of new protocols to facilitate practical use. Due to the rise of streaming media and other services and the increasing demand for bandwidth by families and enterprises, the previous generation protocols have been unable to meet the use requirements. Therefore, a new generation protocol 802.11 N was launched in 2009. Based on 2.4GHz, it uses technologies such as multiple input multiple output (MIMO), beamforming and 40MHz binding, which makes the transmission distance longer and the rate up to 600mbps.

Over time, a growing number of protocols use 2.4GHz frequency band for transmission, and the available bandwidth is severely compressed. Therefore, the fifth generation protocol 801.22 AC focuses on 5GHz band optimization. In this generation protocol, the working bandwidth and frequency modulation efficiency of a single channel are improved while maintaining good backward compatibility. In addition, it supports multi-user multiple input multiple output (MU-MIMO). MU-MIMO routing signals can be split in multiple dimensions; Compared with the previous generation MIMO technology, it can realize parallel processing. The introduction of related technologies not only improves the real-time transmission rate, but also optimizes the network resources.

At present, 802.11 ax protocol is the latest version, which will be launched in September 2019. Compared with 802.11 AC. In addition to further optimizing the 5GHz frequency band, the protocol also pays attention to the 2.4GHz frequency band. For 802.11 protocol, new technologies are introduced in each iteration to improve the network performance.

TABLE I. IEEE802.11 PROTOCOL OF EACH GENERATION

Protocol Code	Frequency (Ghz)	Signal	Release Time	Performance Evolution
802.11	2.4	FHSS/DSSS	1997	It is one of the first generation wireless LAN standards
802.11a	5	OFDM	1999	Teee802.11a provides higher speed in the whole coverage range, and the specified frequency point is 5GHz. At present, the frequency band is not used much, and there is less interference and signal contention. 802.11a also adopts CSMA / Ca protocol. However, in the physical layer, 802.11a adopts orthogonal frequency division multiplexing
802.11b	2.4	HR-DSSS	1999	It can be used not only as a supplement to the wired network, but also as an independent network, so that network users can get rid of the constraints of network cables and realize real mobile applications. One of the key technologies of IEEE 802.11b is the use of compensation code keying CCK modulation technology, which can realize dynamic rate conversion.
8002.11g	2.4	OFDM	2003	The mission is to give consideration to 802.11a and 802.11b and pave the road and bridge for the transition from 802.11b to 802.11a. The modulation modes specified in 802.11g include OFDM used in 802.11a and CCK used in 802.11b. By specifying two modulation modes, it not only achieves the data transmission speed of IEEE802.11a 54mbit / s with 2.4GHz frequency band, but also ensures the compatibility with IEEE 802.11b products.
802.11n	2.4/5	OFDM	2009	The theoretical rate can be up to 600mbps. 802.11n can work in two frequency bands of 2.4GHz and 5GHz.
802.11ac	5	OFDM	2013	802.11ac is the successor of 802.11n. It adopts and extends the air interface concept derived from 802.11n, including wider RF bandwidth (up to 160MHz), more MIMO spatial streams (up to 8), multi-user MIMO and higher-order modulation (up to 256qam)
802.11ax	2.4/5	OFDMA	2019	Orthogonal frequency division multiple access, multiuser multiple input multiple output, high-order modulation, target wake-up time

Wi-Fi alliance promotes the development of relevant standards. The predecessor of Wi-Fi alliance is the wireless Ethernet Compatibility Alliance (WECA) [2]. In 1999, in order to promote the formulation of IEEE 802.11b specification, the wireless Ethernet Compatibility Alliance was formed. In addition, the alliance also provides verification services for products that meet relevant standards to solve the compatibility problems between different devices, so as to promote the development of IEEE 802.11 protocol. In 2002, WECA was renamed Wi-Fi alliance. At present, Wi-Fi alliance has named some standards the scale is simplified, in which the latest generation protocol IEEE 802.11 ax is called Wi-Fi 6 [3]; IEEE 802.11 AC is called Wi-Fi-5.

C. Innovation and optimization of Wi-Fi-6

The performance of Wi-Fi-6 has been greatly improved. With the increasing number of application scenarios such as video conferencing and mobile teaching, the number of terminal devices using the network is also rising. The increasing number of terminal devices will affect the network efficiency. At present, Wi-Fi6 [4] introduces uplink MU-MIMO, OFDMA orthogonal frequency division multiple access, 1024-qam high-order modulation and other technologies, which will solve the problems of network capacity and transmission efficiency from the aspects of spectrum resource utilization and multi-user access. The goal is to increase the average throughput of users by at least four times and the number of concurrent users by more than three times compared with today's Wi-Fi-5 in a dense user environment.

TABLE II. COMPARISON OF VARIOUS SPECIFICATIONS OF WI-FI 4-WI-FI 6

	Wi-Fi 4	Wi-Fi 5		Wi-Fi 6
Agreement	802.11n	205.11ac		802.11ax
		Wave 1	Wave 2	
Operating frequency band	2.4/5GHz	5GHz		2.4/5GHz
Maximum bandwidth	40MHz	80MHz	160MHz	160MHz
Maximum modulation	64QAM	256QAM		1024QAM
Single stream bandwidth	150Mbps	433Mbps	867Mbps	1201Mbps
Maximum spatial flow	4X4	8X8		8X8
MU-MIMO	N/A	N/A	down	Up/down

1) Orthogonal frequency division multiple access (OFDMA)

Transition from OFDM to OFDMA. Before Wi-Fi6, data transmission was carried out using OFDM mode. In this mode, a single user will occupy all subcarriers and send a complete packet in a time segment. Although this transmission mode can meet the needs of a single user, a single user does not need to use all subcarriers when the data packet is small. Therefore, this transmission mode will cause a waste of network resources to a certain extent, and will increase the waiting time of other users in the case of multiple users. In order to improve user network experience, OFDMA is introduced into Wi-fi-6 protocol. OFDMA realizes multi-user multiplexing channel resources by allocating subcarriers to different users and adding multiple access in OFDM system. In addition, in Wi-Fi6 protocol, the minimum subchannel "resource unit" (RU) contains at least 26 subcarriers. Since user data is carried on the ru through subcarriers, multiple users can be transmitted at the same time in each time slice [5].

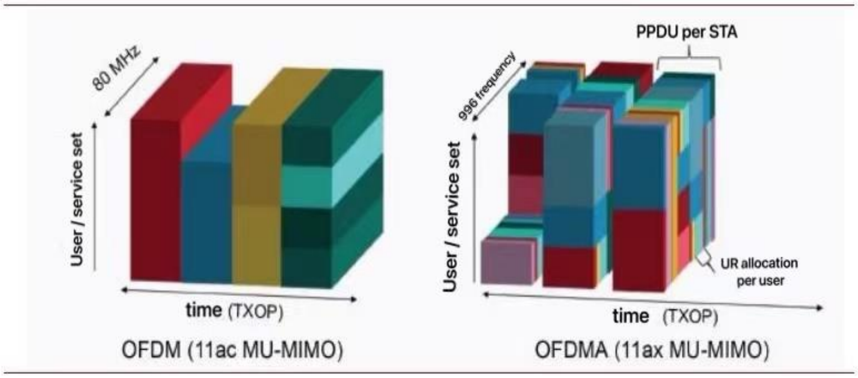


Figure 2. Comparison between OFDM and OFDMA

OFDMA transmission mode can better adapt to small packet usage scenarios. Due to the poor channel state of some nodes in the actual transmission process, if it cannot be adjusted effectively, there is the possibility of data loss. However, this phenomenon can be effectively alleviated in wi-fi6. Since Ru is the smallest sub channel in OFDMA transmission mode and wi-fi6 can allocate transmission power according to channel quality, it can realize transmission using optimal Ru resources. For users, the bandwidth requirements are different in use; In the ofmda mode, a single customer can use one or more groups of Rus to meet the bandwidth requirements. In the multi-user scenario, because multiple users can share the channel in OFDMA transmission, the delay will be effectively improved compared with OFDM. OFDMA transmission mode can meet the different needs of users. It has higher transmission efficiency and better effect in small data packets.

2) Multiuser Multiple Input Multiple Output (MU-MIMO)

MIMO technology improves data throughput. MIMO technology includes spatial diversity and spatial multiplexing. Spatial multiplexing can transmit multiple data of a single user or multiple users at the same time without changing the channel bandwidth. MIMO technology can be divided into single user MIMO (SU-MIMO) and multi-user MIMO (MU-MIMO). In the transmission process of SU-MIMO, the AP can only communicate with one user, which can increase the throughput of a single user. Compared with SU-MIMO, MU-MIMO can transmit with multiple terminals at the same time. Since MU-MIMO technology can realize concurrent transmission between AP and multiple terminals, the data throughput at the same time is improved. In the 802.11ac WAVE2 standard, the MIMO technology introduced only supports data downlink and can only transmit data to four users at most at the same time. The uplink data of users is still transmitted one by one and cannot be concurrent. However, in Wi-Fi6, this technology is more fully utilized.

Wi-Fi-6 uses full MU-MIMO technology. At the data downlink end, some versions of 802.11ac protocol support DL 4x4 MU-MIMO; In Wi-Fi-6, DL MU-MIMO is further improved and supports 8x8 transmission mode. At the data uplink, in the previous protocol, only UL SU-MIMO is supported, while Wi-Fi6 introduces UL MU-MIMO for the first time to transmit data on multiple spatial streams using the same channel resources in the case of multiple users. Therefore, after introducing DL MU-MIMO into Wi-Fi-6, DL / UL MU-MIMO technology has been supported in the protocol. With the support of MU-MIMO technology, the performance of Wi-Fi6 will be improved in the multi-user data transmission environment [6].

Summary:Under the Wi-Fi-6 standard, OFDMA and MU-MIMO develop together. From a technical point of view, OFDMA supports multiple users to improve concurrency efficiency by subdividing channels (subchannels), and MU-MIMO supports multiple users to improve throughput by using different spatial streams. In Wi-Fi-6 protocol, these two technologies can be used at the same time. Based on the cooperative development of different technologies, the transmission speed can be improved while the delay is effectively reduced. In the case of multiple users, the user network experience has been effectively improved.

TABLE III. COMPARISON OF OFDMA AND MU-MIMO TECHNOLOGIES

OFDMA	MU-MOMO
Improve efficiency	Increase capacity
Reduce delay	Higher single user rate
Best for low bandwidth applications	Best for high bandwidth applications
Most suitable for small packet message transmission	Most suitable for large packet message transmission

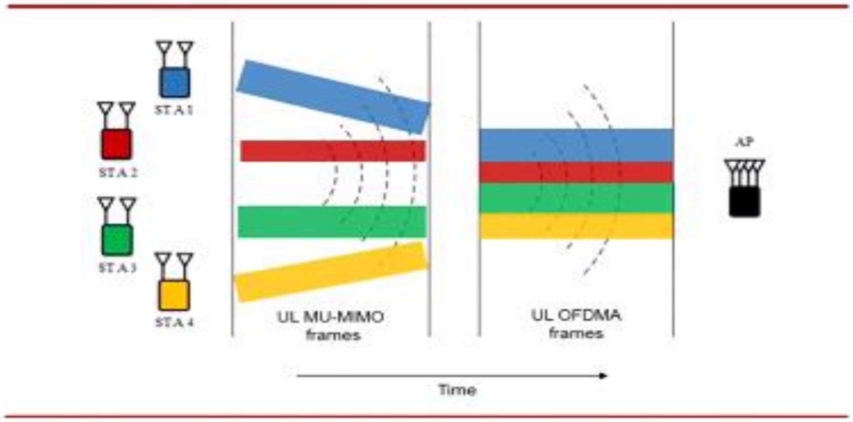


Figure 3. Multi user mode uplink scheduling sequence

3) Target wakeup time

TWT technology can make the terminal have longer endurance. With the development of science and technology, more and more electronic devices join the wireless network. On the consumer side, in addition to mobile phones, notebooks and other electronic devices, there are a large number of smart home devices in the home wireless network. Most of these devices use battery power supply. If they are active and not working for a long time, there will be the problem of power waste. Wi-Fi-6 introduces TWT technology, which allows the device to negotiate the wake-up time and enter the sleep state without data transmission. This technology can effectively reduce battery consumption and achieve longer standby time.

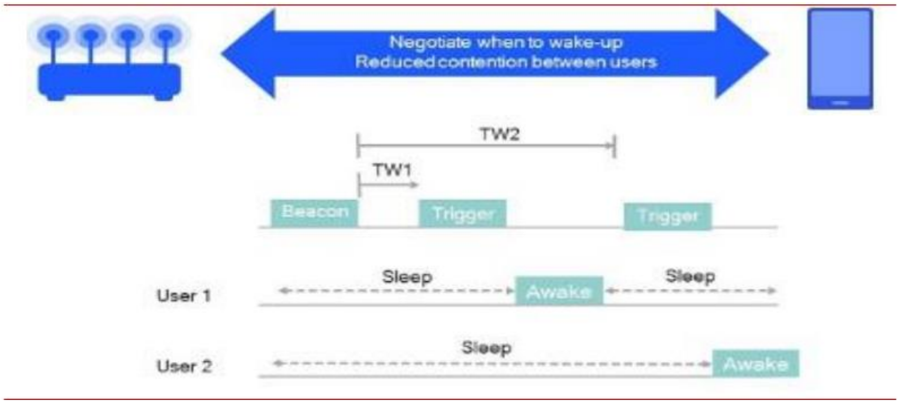


Figure 4. TWS Technology Exhibition

Summary: In addition to the technologies mentioned above, higher adjustment technology and BSS coloring coloring mechanism are also introduced into Wi-Fi-6 in terms of performance improvement. With the support of relevant technologies, the communication between the terminal and the AP is smoother through specific technologies, which not only improves the network carrying capacity, but also reduces the delay. On the consumer side, the penetration rate of smart home devices is increasing, which puts forward higher requirements for network carrying capacity. With the advancement of the Internet of things, the number of networking devices will continue to rise, whether consumer or industrial, and Wi Fi network will become one of the network access options for wireless devices.

II. 5G AND WIFI 6 DEVELOP TOGETHER

A. Competition between Wi-Fi and mobile network

Wi-Fi network has more advantages in use cost. From the spectrum used by Wi-Fi-6 and 5g, the use of Wi-Fi network is unauthorized spectrum, and the relevant spectrum can be used for data transmission when ensuring that the use right of others is guaranteed. However, for mobile communication spectrum, most countries in the world authorize the use in the form of auction, so 5g spectrum is the same as other mobile communication spectrum, and operators obtain the use right through auction. In addition, Wi-Fi network is an extension of wired network, which is more based on fixed network. For consumers, these differences are mainly reflected in the use cost [7]. When consumers use the mobile network, they pay according to their usage. Although Chinese consumers do not need to bear the cost of spectrum licensing, consumers in other regions need to share the relevant costs. Because Wi-Fi networks are more based on fixed networks and can access several terminals, when the fixed network cost is relatively fixed, the cost of a single terminal will decrease with the increase of access quantity. Therefore, considering the use cost of both, Wi-Fi network will have more advantages.

China's Wi-Fi penetration has reached a high level, and the improvement of fixed network performance is conducive to the user experience. According to quest mobile data, the penetration rate of Wi-Fi in mobile phones of mobile Internet users in China continues to rise and has been close to 90%. For users, there may be some dependence on Wi-Fi networks, so as to form usage habits. In terms of network construction, according to the 2019 communication industry statistical bulletin issued by the Ministry of industry and information technology, by the end of December, the total number of fixed Internet broadband access users of the three basic telecom enterprises had reached 449 million, with a net increase of 41.9 million in the whole year. In addition, the number of Internet broadband access ports in China reached 916 million, a net increase of 48.26 million over the end of last year. Among them, the number of FTTH / 0 ports increased by 64.79 million over the end of the previous year, reaching 836 million, accounting for 91.3% of the Internet access ports from 88.9% at the end of the previous year; By the end of June 2020, the total number of fixed Internet broadband access users of the three basic telecom enterprises had reached 465 million, a year-on-year increase of 7%, a net increase of 15.73 million over the end of the previous year. Among them, there are 434 million FTTH / O users, accounting for 93.2% of the total fixed Internet broadband access users. According to the published data, China's fixed network has a high popularity, and under the promotion of national policies such as "copper retreat and optical advance", the optical fiber penetration has reached a high position. Fiber into the home will improve the network carrying capacity. For the Wi-Fi network based on the fixed network, with the support of the equipment, the Wi-Fi network performance will also be improved, making the user experience better.

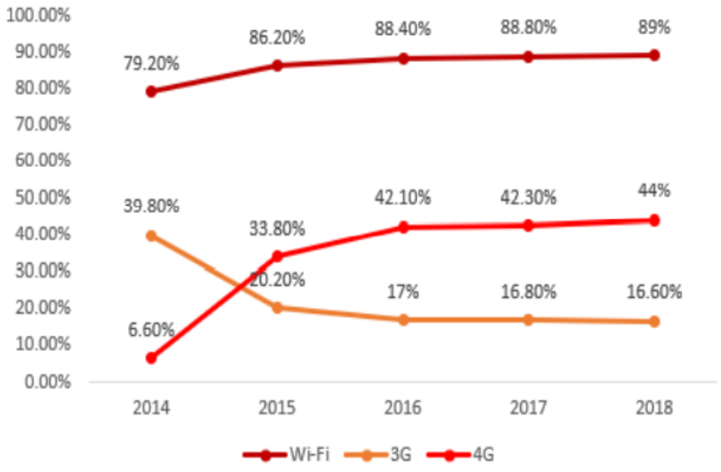


Figure 5. Penetration of mobile Internet users in different network environments;

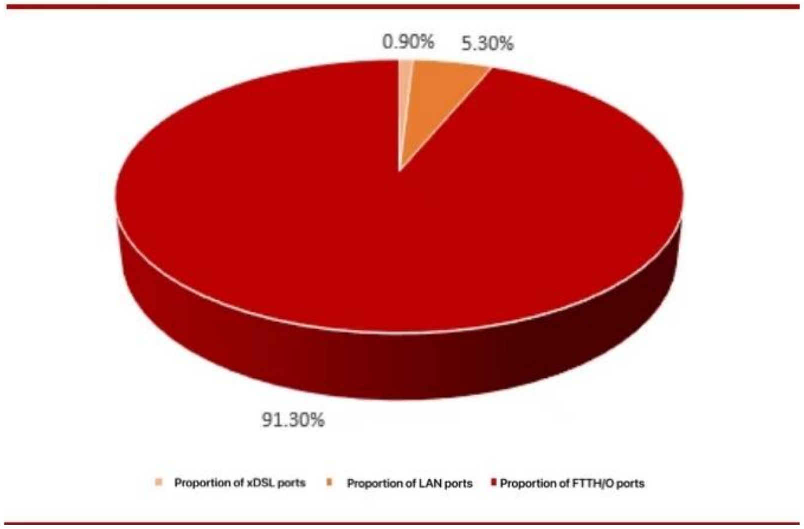


Figure 6. Proportion of Internet broadband access ports in China 2019

Summary: The statement that mobile network will replace Wi-Fi will cause heated discussion in the promotion of new mobile communication technologies of each generation. At present, 5g commercial has been realized in leading countries, and the performance of 5g communication technology has been greatly improved. However, for Wi-Fi networks, relevant technical standards are also constantly updated, and more technologies are introduced to improve network performance and provide users with a better use experience. In addition, from the perspective of the iteration cycle of relevant technologies, the update cycle of mobile communication technology is about 10 years, while the Wi-Fi standard is iterated every 5 years. When the Wi-Fi usage rate is relatively low and the superposition user usage habit has been formed, the possibility of Wi-Fi network being replaced is low.

B. Coordinated development of Wi-Fi-6 and 5G

There is a gap between Wi-Fi-6 and 5g in performance, but each has its own advantages. 5g and wi-fi6 are the next generation mobile wireless technology and the next generation Wi-Fi technology respectively. Although the two network architectures are different, they have a great improvement in performance compared with the previous generation. Among them, 5g communication bandwidth has been greatly improved compared with 4G communication, and the theoretical downlink speed is up to 10 GB / s. In addition, the delay of 5g communication can reach the millisecond level, and the connection density can reach 1 million networked devices per square kilometer. Both the delay and connection density are 10 times higher than that of 4G communication. Therefore, the three downstream application scenarios of 5g can be realized: eMBB (enhanced mobile broadband), uRLLC (highly reliable and low delay connection) and mMTC (massive IOT). In contrast, Wi-Fi-6, according to the 2019 intelligent hardware quality report (phase II) released by China Mobile [7], it selects different standard routers of the same brand for testing. The test results show that Wi-Fi-6 routers have significantly improved performance in terms of transmission rate and delay. Moreover, with the increase of the number of users, it becomes more and more obvious. In 5g band, compared with Wi-Fi-5, the rate and delay of single user are increased by 29% and reduced by more than 5ms respectively; In the case of multiple users, the rate is increased by more than 47%. Overall, the fastest downlink speed of wi-fi6 is 9.6gb/s, and the average network delay is reduced to 20ms. Because Wi-Fi networks and 5g networks use different architectures, there are differences in performance, but they also have different advantages.

5G and Wi-Fi-6 are applicable to different scenarios, and the two will develop together. Although both 5g and Wi-Fi-6 can achieve high-density wireless access and high-capacity wireless services [8], the emphasis between them will be different. 5g communication uses sub 6 GHz high-frequency part or even millimeter wave for transmission. Although using higher frequency transmission can improve the transmission capacity, it also makes the electromagnetic wave length shorter, the signal coverage smaller than 4G communication, and the signal penetration ability decreases. Therefore, 5g will have better performance in open places outdoors. In addition, because 5g mobile network has continuous coverage capability, it can provide service support for mobile terminals, which is an advantage that any short-range communication technology does not have. Therefore, 5g has unique advantages in outdoor scenes, such as Internet of vehicles. In the indoor part, although the 5g signal micro site or pico site can effectively improve the indoor 5g signal coverage, consumers also have a more economical demand for high-speed Internet access in relatively fixed environments, such as homes and offices, and Wi-Fi network can meet the corresponding requirements. With the increasing number of indoor intelligent devices, the demand for high-density connections continues to increase. In addition, HD video, VR and voice calls are very sensitive to bandwidth and delay. Therefore, the advantages of Wi-Fi-6 are highlighted in these scenarios. In addition, if all networking devices used indoors are added with SIM module, it will have a certain impact on product design, product price and convenience of consumers. Therefore, 5g and Wi-Fi-6 have comparative advantages in different scenarios, and they will develop together [9].

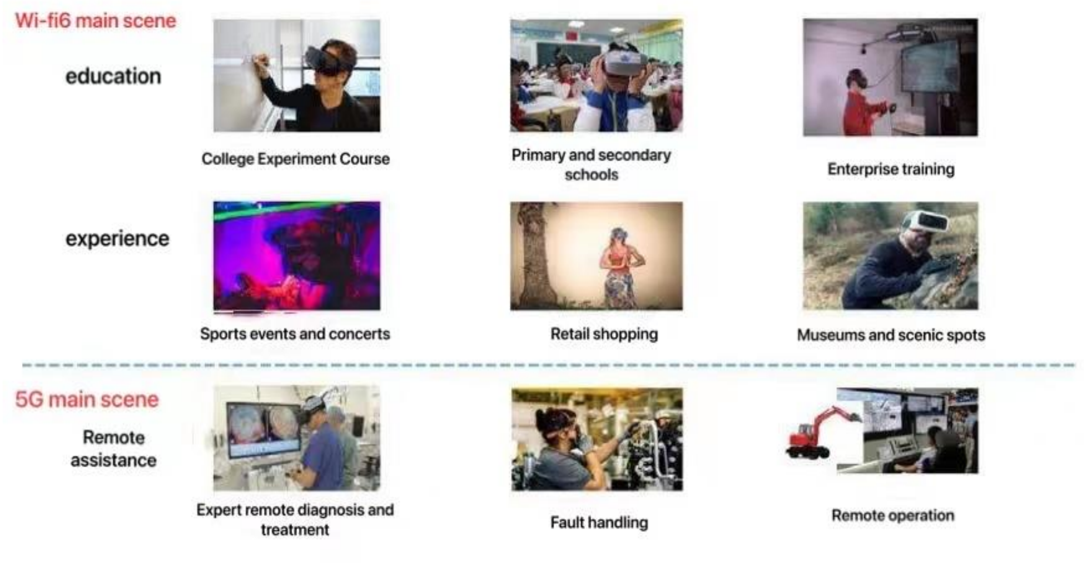


Figure 7. 5G and Wi-Fi-6 usage scenarios

Summary: 5G communication and Wi-Fi-6 have greatly improved performance compared with their previous generation standards. Both can achieve high-density wireless access and high-capacity wireless services. However, due to different construction frameworks, they can adapt to different use scenarios. Among them, 5g communication outdoor space has advantages and can provide communication support for high-speed mobile terminals, while Wi-Fi-6 can provide more economical network services for the Internet of things indoors. Therefore, both can provide corresponding network support for Internet of things services, but the emphasis is different, and the two will develop together.

III. HIGH QUALITY NETWORK CONNECTION AND DEVELOPMENT OF INTERNET OF THINGS DRIVE WI-FI-6 DEMAND

The demand for network use is an important thrust driving the development of Wi-Fi-6. At the beginning of Wi-Fi-6 design, high density and high capacity network service support, including outdoor large public places, high-density venues, indoor high-density wireless office, electronic classrooms and other scenarios. Because 5g network has more advantages in outdoor environment, indoor is an important application scenario of Wi-Fi-6. In indoor space, the demand for information consumption and entertainment, smart device connection and office has become an important driving force driving the development of Wi-Fi-6.

A. *prominent demand for high-quality network connection*

4K / 8K UHD video has increased bandwidth requirements. Ultra high definition video technology is a new round of major technological innovation in the video industry after digitization and high definition, which is of great significance to consumers and the cultural industry. For the development of related industries, China's Ministry of industry and information technology and other departments have issued the action plan for the development of UHD video industry (2019-2022), which proposes to continuously promote the construction of 4K UHD TV content and enrich the effective supply of program content.

Create a number of typical applications of ultra-high definition video in the fields of radio and television, culture, education and entertainment, security monitoring, medical and health care, intelligent transportation, industrial manufacturing and so on. For consumers, content is the main object of their consumption. Under the promotion of national policies, it will help to speed up the construction of high-definition video content. While the UHD video content is gradually enriched, video transmission has become the focus of attention. In terms of the required bandwidth, the traditional HD service can meet the use requirements by only 20MBps bandwidth. The full 4K requires a bandwidth of more than 100Mbps, while 8K video requires a higher bandwidth. Therefore, ultra-high definition video will put forward higher requirements for transmission bandwidth.

Transmission quality will affect VR user experience. VR is a virtual environment constructed by computer. On the one hand, it can meet the needs of entertainment consumption in 2C scene; On the other hand, the 2B terminal can also inject power into the development of the industry. At present, VR technology is gradually implemented, and some related products have been launched into the market. However, in actual use, due to network transmission and other problems, users will feel uncomfortable and affect the user experience. In addition, with the maturity of cloud computing and edge computing, local VR will also change to cloud VR, and the network requirements will be higher. According to the data released by HUAWEI, VR technology has high requirements for bandwidth and delay. Therefore, with the gradual popularization of VR technology, the demand for network is also gradually increasing.

Summary: from the perspective of relevant business scenarios, UHD video, VR and other applications mostly occur in indoor scenarios. As the main carrier of indoor transmission, Wi-Fi will increase the network load when the demand for relevant services increases. Therefore, the terminal demand will promote the innovation of Wi-Fi standard to deal with the network. As a new generation standard, Wi-Fi-6 can meet the application needs of ultra-high definition video and VR. With the continuous improvement of terminal penetration, it will be conducive to the popularization of relevant standards.

B. *The Demand for Internet of things is gradually increasing*

Wi-Fi network is an important part of smart device connection. The Internet of things (IOT) is an important trend. Unlike user devices such as laptops, IOT devices need deterministic wireless services, such as polling every 5 milliseconds, otherwise they will shut down or low-power services. Traditionally, these needs have been met through proprietary, niche or operator specific technologies, but with excellent economies of scale and simple it management, enterprise Wi-Fi has increasingly become the preferred indoor Internet of things platform. In order to meet these IOT operational needs, Wi-Fi-6 and its IOT features (such as low power consumption and certainty) are expected to accelerate this adoption. According to Cisco data, Wi-Fi is already the third largest connection mode of the Internet of things. It is expected that by 2021, the number of Internet of things devices connected through Wi-Fi is expected to reach 12 billion.

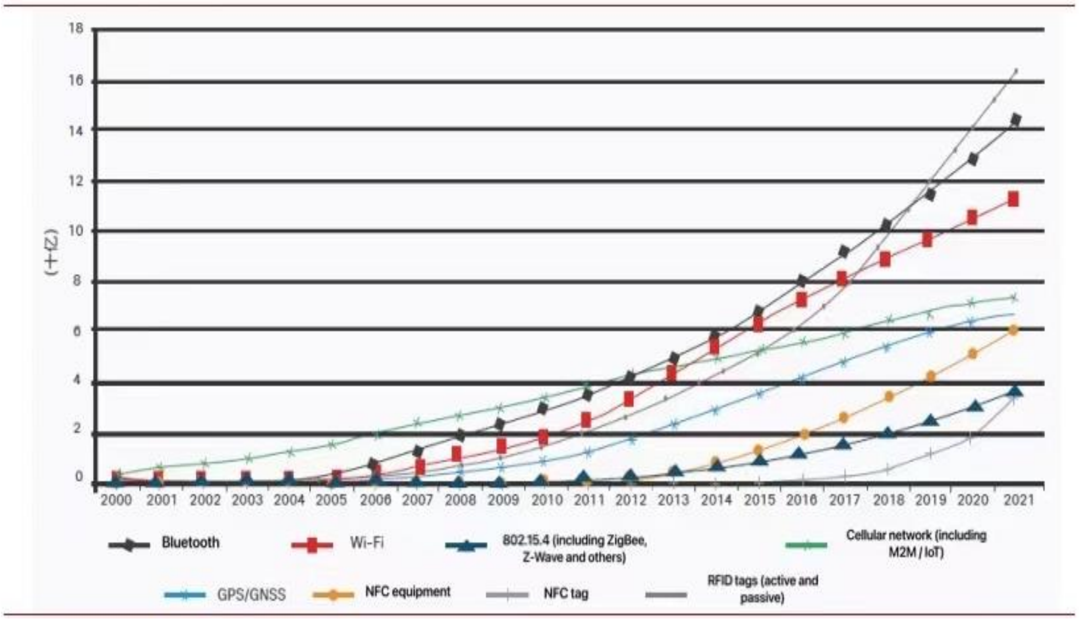


Figure 8. Internet of things trends

1) Wi-Fi-6 enables intelligent industrial production scenarios

In terms of transmission rate, Wi-Fi-6 is upgraded by introducing MU-MIMO technology, and the maximum rate of Wi-Fi-6 can reach 9.6Gbps; 5g communication technology because of the use of mass MIMO technology, the maximum support rate reaches 10Gbps. In terms of transmission capacity, Wi-Fi-6 supports OFDMA technology to enable multiple devices and applications to transmit and receive data at the same time. 5g adopts NOMA technology, which can enable more users to connect without reducing the transmission rate. Although Wi-Fi-6 has similar performance to 5g, Wi-Fi-6 currently has greater advantages in the industrial field [10]. Firstly, the construction of 5g network has just started. It takes a long time from the construction coverage of macro base station to the realization of indoor stable connection of small base stations; However, at present, most factories have generally adopted wired network connection, and it is more convenient to realize Wi-Fi communication connection through wired network. Secondly, the tariff of 5g industrial private network has not been implemented. At present, there are only 5g consumer charges. At present, the initial prices of 5g packages determined by the three communication operators in the first batch are 128 yuan, 129 yuan and 129 yuan respectively, including 30GB / month 5g traffic; However, there are many data to be transmitted in industrial scenarios, and 30GB traffic may not be enough. Wi-Fi-6 connects to a wired network through a wireless router. Wired network charges are fixed charges, while routers can be understood as one-time charges. With the increase of service time, the cost gradually decreases. Therefore, wi-fi6 has cost advantages in industrial scenario applications.

Mettis Aerospace has a huge production base in West Midlands, UK, covering 27 acres, with 515 employees and 3000 manufacturing equipment. In 2019, the company conducted a test on Wi-Fi 6 in the factory. In the test, Wi-Fi-6 uses a channel with a bandwidth of 80 MHz to achieve a download rate of 700 Mbps and a delay of less than 6 milliseconds. The applications implemented on site include 4K video upload, large file transmission, message transmission, voice and video communication, Internet of things sensors, etc. from the test results, Wi-Fi-6 realizes wireless communication with high reliability, high quality, high bandwidth and low delay. Mettis Aerospace plans to deploy Wi-Fi6 throughout the plant as part of a comprehensive IT infrastructure upgrade in the next five years.

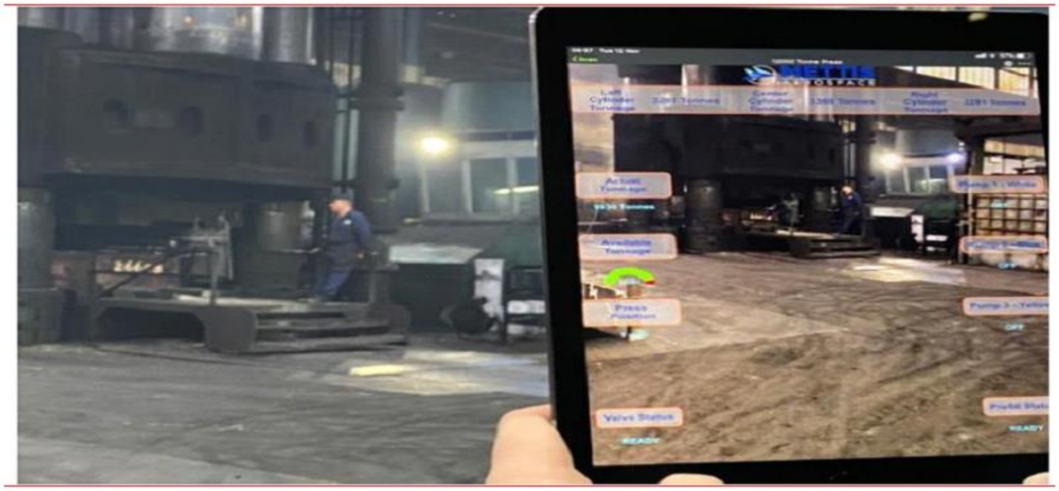


Figure 9. The use of wifi6 was tested in the factory

2) Wi-Fi-6 becomes a smart home network link

Smart home takes residence as the carrier, integrates automatic control technology, computer technology and Internet of things technology, organically combines the functions of home appliance control, environmental monitoring, information management and audio-visual entertainment, and provides a more portable, comfortable, safe and energy-saving family living environment through the centralized management of home equipment. With the development of science and technology, smart home has gradually entered the family life of ordinary residents, and will become a part of family life in the future. According to strategy analytics, the global smart home market reached US $84 billion in 2017, an increase of 16% over US $72 billion in 2016. In 2018, the total consumer spending on global smart home devices, systems and services will be close to US $96 billion, and the CAGR will reach 10% to US $155 billion in the forecast period (2018-2023). North America will account for 41% or $40 billion of total expenditure, followed by the Asia Pacific region of $26 billion and Western Europe of $17 billion.

In China, according to the statistical data of the report on market prospect and investment strategy planning of China's smart home equipment industry, the scale of China's smart home market in 2015 was 40.34 billion yuan, a year-on-year increase of 41.0%. The report shows that the scale of China's smart home market is expected to reach 142.2 billion yuan in 2019, showing a trend of increasing by 100 billion yuan year by year. If the average annual compound growth rate is about 38.13% in the next five years (2019-2023), we predict that the scale of China's smart home market will reach more than 200 billion yuan in 2020 and exceed 500 billion yuan in 2023.

There are three mainstream protocols for smart home: Bluetooth, Wi-Fi and ZigBee. Bluetooth technology is an open global standard for wireless data and voice communication. It is a special short-range wireless technology connection based on low-cost short-range wireless connection to establish a communication environment for fixed and mobile devices. ZigBee is a low-speed and short-distance wireless network protocol. The bottom layer is the media access layer and physical layer based on IEEE 802.15.4 standard. The main features are low speed, low power consumption, low cost, support for a large number of online nodes, support for a variety of online topologies, low complexity, fast, reliable and safe. Although the three protocols are short-range connection protocols, there are differences in performance. From the transmission distance, Wi-Fi > ZigBee > Bluetooth.In terms of power consumption, Wi-Fi > Bluetooth > ZigBee, and Bluetooth and ZigBee devices can be powered by batteries.In terms of transmission rate, Wi-Fi > ZigBee > Bluetooth.

TABLE IV. COMPARISON OF THREE INTERNET OF THINGS PROTOCOLS

	Bluetooth	**Wi-Fi**	**ZigBee**
Network organization	Dot communication network	Star Communication Network	Mesh communication network
Maximum transmission rate	2 Mpbs	300 Mpbs	250Mpbs
Transmission range	10-100m	100-300m	50-300m
Power consumption	low power consumption	High power consumption	secondary

According to the data of the National Bureau of statistics, the per capita residential construction area in China's cities in 2016 was 36.60 square meters. According to the calculation of a family of four, the living area of a family is about 144 square meters. If the family wants to arrange the smart home system, the transmission distance of Bluetooth is not enough to support, and Wi Fi and ZigBee will be better choices. Compared with Wi Fi, ZigBee has advantages in transmission range and power consumption when the maximum transmission rate is slightly inferior. However, in terms of products, at present, ZigBee technology mainly adopts 2.4GHz of ISM frequency band in China, which has weak diffraction ability and wall penetration ability, is vulnerable to obstacles, and is vulnerable to interference from Wi Fi and Bluetooth in the same frequency band. In addition, the development of main products is difficult, the development cycle is long, the product cost is high and the penetration rate is low. Therefore, at present, Wi Fi is the mainstream use protocol of smart home.

IV. WI-FI6 MARKET IS EXPANDING RAPIDLY, AND THE INDUSTRY IS DISTRIBUTED FROM TOP TO BOTTOM

A. Rapid expansion of Wi-Fi-6 market scale

According to the data of Cisco white paper, the number of M2M connections will reach 14.7 billion and CAGR will reach 19% in 2023. In addition, the number of M2M connections will be half of all connections. In M2M applications, home connection will become a common phenomenon. By 2023, home connections will account for 48% or nearly half of the total M2M connections. For a long time, one of the main solutions to meet the growing bandwidth demand is to use Wi-Fi network to enable operators to expand capacity to meet the needs of their users. By 2023, there will be nearly 628 million public Wi-Fi hotspots in the world, up from 169 million hotspots in 2018, an increase of four times. Regionally, the compound annual growth rate in central and Eastern Europe reached 38%, the compound annual growth rate in Asia Pacific and Latin America reached 37%, the compound annual growth rate in the Middle East and Africa reached 30%, and the compound annual growth rates in North America and Western Europe were 25% and 20% respectively. According to IDC data, the overall scale of WLAN market reached US $230 million in the third quarter of 2019, which is in a steady upward trend [11].

Although the Wi-Fi Alliance announced the Wi-Fi-6 standard in October 2018, the relevant certification plan was officially launched on September 16, 2019. After the release of relevant standards, although some manufacturers have launched relevant routing products, there are some disadvantages, such as less terminal product support and higher product pricing. After the certification work is started, manufacturers will be more active in the introduction of relevant technologies, so as to promote relevant product development. The promotion of Wi-Fi-6 certification is conducive to the expansion of Wi-Fi-6 market scale. IDC predicts that in 2020, the scale of Wi-Fi-6 market in China will be close to US $200 million, and by 2023, the scale of China's Wi-Fi-6 market will reach US $1 billion, with an annual compound growth rate of 71%.

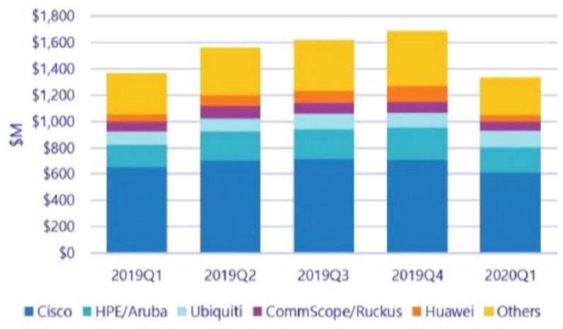


Figure 10. 2019Q1-2020Q1 enterprise router market share

B. Wi-Fi receiving terminal: Mobile flagship covers the whole line, followed by tablet

1) The flagship products of mobile phones widely support Wi-Fi-6

Major mobile phone manufacturers have released Wi-Fi-6 standard mobile phones, and their penetration in mobile phones will gradually increase. Before the Wi-Fi alliance started Wi-Fi-6 standard certification, only a few mobile phone manufacturers launched corresponding terminals. Take Samsung as an example. In February 2019, the new generation of flagship Galaxy S10 series mobile phones was officially released. Since no manufacturer has released relevant products before, Galaxy S10 series mobile phones will become the first batch of mobile phones in the world to support Wi Fi 6. Since then, Samsung has also introduced relevant technologies into the note 10 series. However, not all mobile phones support Wi-Fi-6, and only note 10 + models provide relevant support. Although Samsung launched Wi-Fi-6 standard mobile phones in early 2019, they were not introduced in all subsequent mobile phones. Most mobile phone manufacturers have not launched relevant products. However, this situation has improved. From the mobile phones released by various manufacturers, except Samsung and apple, most domestic mobile phone manufacturers will launch Wi-Fi-6 products in 2020. At present, all mainstream brands are beginning to pay attention to the Wi-Fi-6 standard. Although most mobile phones supporting Wi-Fi-6 standard are brand flagship models with relatively high price, with the development of Wi-Fi-6 certification and increasing attention, more mobile phones will support Wi-Fi-6 standard and their penetration rate will gradually increase.

2) Tablet devices follow closely

Compared with mobile phones, tablet computers have the same functions as mobile phones except for the lack of mobile phone functions. Tablet computers have larger battery capacity, wider screen vision and higher definition display. If the tablet computer is equipped with touch pen, external keyboard and other accessories, it can undertake light office tasks. Tablet has been recognized by consumers because of its large screen, portability, office and other characteristics. Tablet computers have become one of the electronic devices that consumers carry when they travel. Affected by the epidemic in 2020, home office and distance education have become measures taken by many countries and regions in the world at the time of serious epidemic. In order to adapt to telecommuting and distance teaching, the sales of tablet computers have increased greatly. According to Canalys data, the sales volume in the second quarter of 2020 reached 37.5 million, a year-on-year increase of 26.1%. Among them, Apple's sales increased by 19.8% year-on-year, accounting for 38.0% of the total sales; Samsung's sales increased by 39.2% year-on-year, accounting for 18.7% of the total sales. HUAWEI's sales increased by 44.5% year-on-year, accounting for 12.7% of the total sales.

TABLE V. Q2 TABLET PC SHIPMENTS IN 2020

Vender(company)	Q2 2020 Shipments	Q2 2020 Marker Share	Q2 2019 Shipments	Q2 2019 Marker Share	Annual Growth
Apple	14,249,000	38.0%	11,894,000	40.0%	19.8%
Samsung	7,024,000	18.7%	5,048,000	17.0%	39.2%
Huawei	4,770,000	12.7%	3,300,000	11.1%	44.5%
Amazon	3,164,000	8.4%	2,308,000	7.8%	37.1%
Lenovo	2,810,000	7.5%	1,838,000	6.2%	52.9%
Others	5,525,000	14.7%	5,379,000	18.1%	2.7%
Total	37,542,000	100.0%	29,767,000	100.0%	26.1%

Although the tablet has released LTE version or cellular version, cellular data can be used through SIM card. However, from the use environment, tablet computers are mainly used indoors. In the indoor environment, cellular communication may be easily interfered or intercepted, and the use experience needs to be improved. In addition, considering that the scenarios when using tablet computers are generally video or games and other applications that consume a lot of traffic, if cellular data is used, it may cause a large tariff burden. Therefore, tablets generally connect to the network through Wi-Fi. As a new generation of wireless WLAN technology, Wi-Fi-6 has the advantages of low delay, high capacity and high rate. It has become a new element for tablet manufacturers to release new products. Taking the top three tablet computer manufacturers as an example, HUAWEI, SAMSUNG and APPLE have launched their own tablets supporting Wi-Fi-6 in 2020.

C. Routing equipment: domestic brands catch up and split the development trend of key fields

1) Overseas giants have obvious advantages, and domestic brands catch up

The terminal is responsible for receiving the network signal, and the router is responsible for transmitting the network signal. Router is a hardware device connecting two or more networks. It acts as a gateway between networks. It is a special intelligent network device that reads the address in each data packet and then determines how to transmit it. It can understand different protocols, such as Ethernet protocol used in a LAN and TCP / IP protocol used in the Internet. In this way, the router can analyze the destination address of data packets transmitted from various types of networks, convert the address of non TCP / IP network into TCP / IP address, or vice versa. Then, according to the selected routing algorithm, each packet is transmitted to the specified location according to the best route. So routers can connect non TCP / IP networks to the Internet.

According to the functional classification, routers can be divided into backbone routers, enterprise routers and access routers. Backbone routers realize the interconnection of enterprise networks. The requirements for it are speed and reliability, while the cost is secondary. In 2018, the annual revenue of HUAWEI routers in the operator market increased by 8.6%, ranking first in the list with 30% market share. This is also the first time that HUAWEI's IP core router has surpassed the annual market share of the overall operator's IP router field after ranking first in the global operator market share in 2017. In 2019, HUAWEI's router products ranked first in the market share of global operators for three consecutive years. Enterprise or campus level routers connect many terminal systems. Its main goal is to realize as many endpoint interconnections as possible at the lowest cost, and further require to support different quality of service. According to IDC data, Cisco's WLAN revenue fell 6.7% year-on-year to $611 million in Q1 2020. Cisco remains the market share leader, with a market share of 45.7% in the quarter, up from 44.6% in 2019. HPE Aruba's revenue increased by 14.2% year-on-year, and its market share increased from 13.8% in 2019 to 14.4% in the first quarter of 2020. Ubiquiti's WLAN revenue increased by 24.8% year-on-year, with a market share of 9.5%, higher than 7.0% in 2019. CommScope (formerly arris / ruckus) revenue fell 4.7% year-on-year and its market share was 5.2%. HUAWEI's revenue fell 15.0% year-on-year and its market share was 3.8%. According to IDC data, Ruijie ranks second in China's enterprise WLAN market with a market share of 23.95%. Among them, Ruijie ranks first with 40.66% in the Wi-Fi-6 category market.

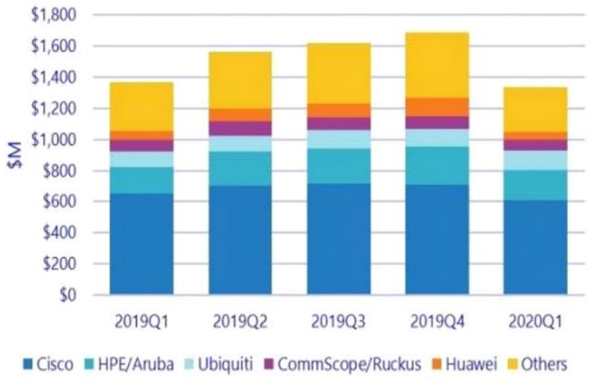


Figure 11. 2019Q1-2020Q1 enterprise router market share

Access routers connect small business customers in homes or ISPs. Access level routers are divided into single frequency routers and dual frequency routers. The single frequency router is a traditional router, and the Wi-Fi signal is only 2.4GHz. 2.4GHz is a low-frequency signal, which has the advantages of strong penetration and longer propagation distance. For large houses, it can have better coverage capacity; However, the disadvantage is that most devices are applicable to the 2.4GHz frequency band, and the interference is large where there are many users. In addition to a 2.4GHz Wi-Fi signal, the dual band router also has a 5GHz Wi-Fi signal. Advantages of 5GHz Wi-Fi: wide signal bandwidth, clean wireless environment, less interference, stable network speed, and can support higher wireless rate; The disadvantage is that the attenuation is large when propagating in air or obstacles, and the coverage distance is generally smaller than 2.4GHz.

In terms of brand attention, in 2019, tp link, which has long been at the forefront of the wireless router market, fell in market attention, ranking second with 14.9%. In the past two years, HUAWEI has continued to receive market attention, reaching 16.8% in 2019, ranking first in brand attention.

2) Split key components of Wi-Fi equipment and grasp the development trend

According to the research report forecast of global Wi-Fi chip market scale in 2022 released by market and markets, the global Wi-Fi chip market scale reached US $15.89 billion in 2016 and is expected to increase to US $19.72 billion in 2022. At present, Wi-Fi devices are still dominated by Wi-Fi-5 products, and Wi-Fi-6 products are expected to enter a rapid penetration period in 2020. According to Dell'Oro's prediction, the shipments of chips supporting Wi-Fi-6 will account for 10% of the total shipments in 2019, and will reach about 90% by 2023, becoming a real mainstream product. At present, the mainstream Wi-Fi-6 router CPU suppliers in the market include Hisilicon, MediaTek, Broadcom, Qualcomm, Intel, etc. Most of them adopt Cortex-A53 of arm company, and some also adopt MIPS architecture. In terms of technology, Qualcomm's Wi-Fi 6 chip adopts 14nm process technology, which is relatively advanced. Most manufacturers use 28nm process chips. In terms of main frequency, the minimum specification is dual core processor, Qualcomm Technology is relatively advanced, and the whole system realizes four core processor.

TABLE VI. SOME WIFI 6 CHIP SUPPLIERS AND THEIR PRODUCTS

Manufacturer	Model	Framework	Process Nm	Dominant Frequency	Representative Model
HiSilicon	Hi5651L	Cortex-A53	28	Binuclear 1.2GHz	HuaweiAX3
	Hi5651T	Cortex-A53	28	Tetranuclear 1.4GHz	HuaweiAX3 Pro
MTK	MT7621DAT	MIPS32 1004Kc	28	Binuclear 880MHz	TL-XDR1860
	MT7622B	Cortex-A53	28	Binuclear 1.35GHz	TL-XDR3230
Broadcom	BCM6750	Cortex-A7	28	Trinuclear 1.5GHz	ASUS AX3000
	BCM6755	Cortex-A7	28	Tetranuclear 1.5GHz	ASUS AX56U
	BCM4906	Cortex-A53	28	Binuclear 1.8GHz	ASUS AX92U
	BCM4908	Cortex-A53	28	Tetranuclear 1.8GHz	ASUS AX88U
Qual comm	IPQ6000	Cortex-A53	28	Tetranuclear 1.2GHz	XIAOMI AX1800
	IPQ8071A	Cortex-A53	14	Tetranuclear 1GHz	XIAOMI AX3600
	IPQ8072A	Cortex-A53	14	Tetranuclear 2GHz	ASUS AX89X
	IPQ8074(A)	Cortex-A53	14		
Intel	GRX350	MIPS32 34Kc	40	Binuclear 800MHz	NETGEAR RAX40

An independent wireless chip processes a single signal. It can be divided into two chips responsible for 2.4G signal and 5G signal, or it can be concentrated into one chip. Some models of processors can also process single frequency signals (i.e. 2.4G or 5G) or dual frequency signals at the same time. According to our data statistics, at present, most of them use a single chip to be responsible for a single signal, so there will be two or more wireless chips in the Wi-Fi device. Taking Xiaomi AX3600 router as an example, the IPQ8071A processor is adopted, the chip of 2.4G signal is QCN5024, and the chip of 5G signal is QCN5054.

FEM has become the mainstream of wireless power amplifier. There are three main ways of wireless power amplifier: PA and LNA are encapsulated to form two chips respectively, PA + LNA is integrated, and FEM is adopted. Based on the integration of PA and LNA, FEM adds additional functions such as power detection. According to our statistics, nearly 70% of the power amplifiers in Wi-Fi-6 routers adopt FEM mode.

Some players of MU-MIMO began to break through gradually. Traditional router MU-MIMO is mainly in 2, 3 and 4 modes, and even some products do not have MU-MIMO. In the initial stage of Wi-Fi-6, MU-MIMO of Wi-Fi devices mainly exists in 2 and 4 modes. At present, some players of the router have developed products supporting 4*2 MU-MIMO. For example, ASUS AX11000 router is equipped with 4*2 MU-MIMO. Not only ASUS, but also router brands such as NetWare RAX200, Orbi RBK852 and Velop MX5300 are equipped with 4*2 MU-MIMO.

V. Expectation

The high network speed under the Wi-Fi-6 standard will have a significant impact on our life. Although 5g has many similarities with its application scenarios, Wi-Fi-6 still has its unique role. Such as enterprise WLAN, industrial scenarios such as smart factory and unmanned storage; High density scenes, such as airports, hotels, large venues, etc; New intelligent terminals, such as wearable devices, smart home, Ultra HD applications, VR / AR, etc.; Service scenarios, such as telemedicine, require high speed, large capacity and low delay.

In the future digital construction, Wi-Fi still has long-term vitality. Under the trend of the integration of cloud management, innovation, operation and maintenance and IOT, it is constantly realizing technology evolution and product updating. The value Wi-Fi-6 brings to enterprises is mainly reflected in extensive connection, extreme experience, improving efficiency, optimizing process, innovating business, etc.

In short, compared with the cost of 5g, Wi-Fi-6 can also be regarded as a good complement of 5g, which not only saves cost, but also helps enterprises achieve a high performance level. In the new infrastructure era, the challenge of connection begins with the business needs of the actual scene, such as the demand for extreme speed brought by the intelligent era, the demand for ultra-low delay in emerging fields such as automatic driving, the demand for massive connection, the demand for business integration, and so on.

Eventually, more Wi-Fi- 6 products will spring up.

Reference

[1] Tavsanoglu Ali, Briso César, CarmenaCabanillas Diego, Arancibia Rafael B.. Concepts of Hyperloop Wireless Communication at 1200 km/h: 5G, Wi-Fi, Propagation, Doppler and Handover [J]. Energies, 2021, 14(4).

[2] Anonymous. Wideband Transceivers for 5G, Wi-Fi, UWB Test [J]. Microwave Journal, 2020, 63(12).

[3] Advantech; Advantech Launches Edge Network Appliance Designed Ready for 5G & Wi-Fi 6[J]. Network Business Weekly, 2020.

[4] Advantech Launches Edge Network Appliance for 5G & Wi-Fi 6 [J]. Telecomworldwire, 2020.

[5] Williams Idongesit. Community Based Networks and 5G Wi-Fi[J]. Ekonomiczne Problemy Usług, 2018, 131.

[6] Shensheng Tang, John O'Rourke, Grace Tang. Traffic modelling of an integrated 5G/Wi-Fi network with generally distributed user-dwell times [J]. International Journal of Wireless and Mobile Computing, 2020, 18(3).

[7] Williams Idongesit. Community Based Networks and 5G Wi-Fi [J]. Ekonomiczne Problemy Usług, 2018, 131.

[8] 5G & Wi-Fi Market Study: Provider Strategies for the Next-Generation Network [J]. M2 Presswire, 2016.

[9] Zeus Kerravala. Next-gen wireless options: Wi-Fi 6, 5G or private 5G? [J]. Network World (Online), 2021.

[10] Guidehouse Insights Reports Global Investment in Industrial Wi-Fi 6 Infrastructure [J]. Wireless News, 2020.

[11] Nokia launches world’s first self-optimizing mesh Wi-Fi 6 solution for CSPs [J]. M2 Presswire, 2020.

Research on Digital Image Watermarking Algorithm in Frequency Domain Based on Matlab

Wu Hejing
East University of Heilongjiang, 150086
E-mail: 499917928@qq.com

***Abstract*—With the rapid development of the Internet, more attention has been paid to information security and copyright issues in the network, and people's copyright awareness has gradually been set up. Based on signal processing and image processing, I have studied embedding digital watermark into DCT domain. Through people's copyright awareness has gradually been set up. The maintenance of copyright has become a hot topic, and the development of digital watermarking has solved this problem for people [1]. Thus, watermarks are embedded in color images. Before I add watermark data, I scramble the watermark data. By contrast, the watermark algorithm embedded in the DC component of the RBG model is robust.**

Based on signal processing and image processing, I have studied embedding digital watermark into DCT domain. Through people's copyright awareness has gradually been set up. The maintenance of copyright has become a hot topic, and the development of digital watermarking has solved this problem for people.. Thus, watermarks are embedded in color images. Before I add watermark data, I scramble the watermark data. By contrast, the watermark algorithm embedded in the DC component of the RBG model is robust.

Keywords-DCT; Digital Watermarking; Color Image

I. INTRODUCTION

In the field of copyright protection, the secret technology of information plays a great role. It can fully protect the copyright of the original author through a series of operations. Digital watermarking technology is the main method of information hiding technology. Through digital watermarking technology, some additional information of multimedia digital products (such as image, video, audio, etc.) can be displayed. This information is often used to indicate the source of the product and declare the copyright. The purpose is to prevent the copyright of multimedia digital products from being infringed, tampered or copied.

Digital watermarking technology involves many disciplines, including signal technology, cryptography and so on. Therefore, it is a complex and challenging technology. Each researcher can conduct extensive and in-depth research on digital watermarking technology according to his own understanding and learning of this technology. It can be said that with the development of digital watermarking technology, there have been many mature and effective algorithms and achieved good results. However, while having fruitful results, we should also see that there are still many difficulties and doubts that need to be overcome

DOI: 10.21307/ijanmc-2021-033

by technicians. For example, many existing watermarking algorithms are not robust enough and need to optimize the algorithm to further improve the performance of the algorithm. With the rapid development of computer technology and the wide popularization of the Internet, multimedia information products will be more abundant in the future, making it easier and faster for people to obtain information. Therefore, digital watermarking technology will have more far-reaching application value.

II. THE MAIN RESEARCH CONTENT OF THIS PAPER

In this study, the basic algorithms and knowledge of information hiding are analyzed, and the research perspective is focused on the fusion and scrambling of digital images. After extensive reading and studying the research conclusions of many scholars, a new color image technology algorithm is proposed. The technical support of this new algorithm is wavelet transform technology and chaos fusion technology, the specific performance of this new algorithm is demonstrated by experiments, and the experimental results show that this method is reasonable. The details are as follows:

1) This paper introduces the hiding technology, including its background, significance, research status at home and abroad and its basic principle model, and popularizes the basic attributes and specific applications of information system.

2) The basic principles and properties of image hiding technology are covered, and its main applications are analyzed.

3) The algorithm types of color image steganography are listed, and the algorithm types in the transform domain are mainly studied and analyzed. Finally, according to the specific comparison of different algorithms, the wavelet domain transform is determined as the main method.

4) This paper mainly analyzes the image scrambling technology, briefly introduces the centralized image scrambling method, and discusses the Arnold transform and its application.

5) This paper analyzes the digital watermarking algorithm in wavelet domain in detail, and focuses on the process of embedding and extraction.

6) The stability of the algorithm is analyzed, including the specific impact test. The stability and anti-interference of the digital watermark are verified by comprehensively analyzing the peak signal-to-noise ratio.

For digital watermarking, no matter which algorithm is adopted, it can not guarantee that the digital watermarking itself is perfect. It can only choose any algorithm according to the actual situation. Generally speaking, the algorithm needs to consider three characteristics, namely imperceptibility, robustness and capacity, as shown in Figure 1. [2]

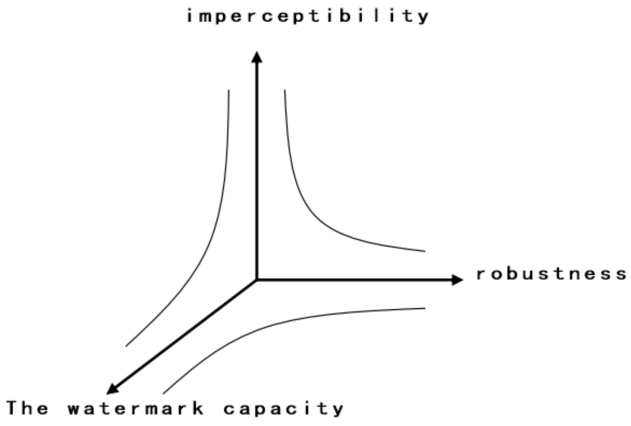


Figure 1. The performance indexes of robust watermarking are reasonable and compromise

III. BASIC MODEL OF DIGITAL WATERMARKING SYSTEM

The basic model of digital watermarking system is mainly the embedding and extraction of watermarking model. In the operation of making watermark model, preprocessing is usually required to ensure that the system can be applied. This preprocessing is to transform the watermark memory in advance. It can be used as the embedded watermark signal only after meeting the requirements. The specific model is shown in the figure. The embedding process of watermark is shown in Figure 2, and the original data represents the difference between blind detection and non blind detection.

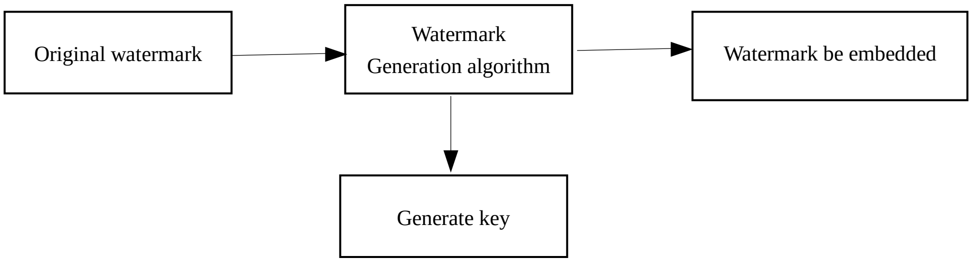


Figure 2. General model of watermark generation

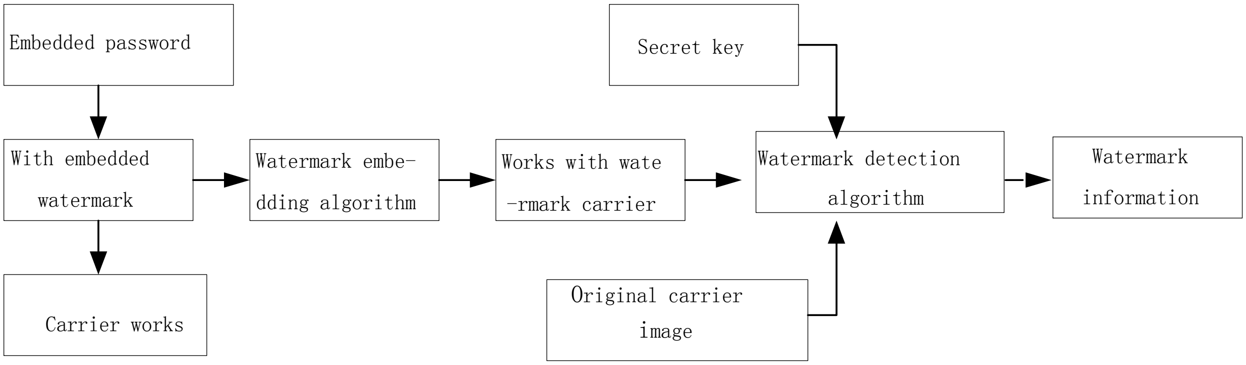


Figure 3. Watermark embedding and extraction process

IV. PERFORMANCE EVALUATION OF DIGITAL WATERMARKING SYSTEM

There are generally two technical means for the evaluation of image quality, one is the subjective evaluation based on human experience or some visual effect, and the other is the objective evaluation based on specific quality indicators such as signal-to-noise ratio. Subjective evaluation will be disturbed by external factors such as people's mood and working state, resulting in inaccurate results. Therefore, objective evaluation plays a very important role in the performance evaluation of digital watermarking system.

The quality evaluation indexes of digital watermarking system are also divided into many categories, which can be divided into variance, signal-to-noise ratio and peak signal-to-noise ratio. Table 2.1 is a specific calculation formula, in which, it represents the pixel size of the color image and the embedded pixel value. The most used indicators by researchers are signal-to-noise ratio and peak signal-to-noise ratio. Error based distortion measurement method:

①Mean square deviation： $MSE = \sum_{m,n}(I_{m,n} - I'_{m,n})^2 / MN$

$$SNR = 10\lg(\sum_{m,n} I^2{}_{m,n} / \sum_{m,n}(I_{m,n} - I'_{m.n})^2)$$

②Signal to noise ratio： $PSNR = 10\lg(MN \max_{m,n} I^2{}_{m,n} / \sum_{m,n}(I_{m,n} - I'_{m,n})^2)$

③Peak signal-to-noise ratio: $IF = 1 - \sum_{m,n}(I_{m,n} - I'_{m,n})^2 / \sum_{m,n} I^2{}_{m,n}$

④Image fidelity：

⑤ Normalization constant： $NC = \sum_{m,n} I_{m,n} I'_{m,n} / \sum_{m,n} I^2{}_{m,n}$

V. FREQUENCY DOMAIN WATERMARKING ALGORITHM

In the current image lossy compression, more concepts are absorbed, such as DFT, DCT and other algorithms are applied to this field. Many watermark robustness algorithms will adopt the above methods. The robustness of watermark is improved through the addition of this method. Moreover, some fragile watermark systems put forward the resistance to lossy compression, which can be easily realized according to the mode of change domain. Therefore, more algorithms can be implemented in the change domain according to the tampered characteristics of the change domain.

Taking Fourier transform as an example, this classical algorithm is also applied to watermark transform algorithm. DFT algorithm can not only realize the non deformation of watermark, but also realize position transformation in the specific embedding process. According to this Fourier transform mode, the amplitude and phase values of some coefficients can be modified to realize watermark embedding. In order to ensure the invisibility and robustness of the watermark, the watermark content can be embedded with different frequency coefficients. O.Ruanaidh [3] did research on related algorithms, two algorithms are obtained: 1. The phase technology of DFT coefficient is mainly used to change the watermark; 2. The change of watermark mainly takes the displacement change of watermark.

Another DCT algorithm in the algorithm is the discrete cosine change algorithm. This transformation method can be for the complete image, or it can divide the image into multiple blocks and divide the image into multiple modules of 8 * 8. In this way, DCT transformation shall be carried out first, and then the embedding space shall be selected according to the mode of the carrier. The stable frequency band shall be selected in the embedding space, and the coefficients of this part of the stable frequency band shall be modified or replaced. The information of the carrier is to reflect its main external structure, which will not lead to too fuzzy pixels. The embedding of digital watermark will not affect the characteristics of the graphics. The high-frequency information is a type other than human perception. The compression technology also removes this part of the content in the compression, and the removed part will also damage the robustness of the watermark. Both high frequency and low frequency are in the range that may affect the watermark, so the best way is to load the information into the if information of the image.

In addition, there is DWT watermarking technology, which is discrete wavelet transform technology. This technology has multi-resolution characteristics. Compared with other transform domain methods, this transform method has better energy concentration characteristics, and after transformation, it is more in line with human visual system. It is also applied more in the new era, including MPEG-4 and JPEG-2000.

Discrete cosine transform, also known as DCT method, is an orthogonal image coding method. If this method was traced back to 1968, Andrews and others were the first to apply it. At that time, he found that the high-frequency component of natural images was in a relatively small amplitude, so it was not important for this part to occupy a position in the whole image system. Therefore, he proposed the transformation coding form of asymmetric mode, Then it is encoded and transmitted according to Fourier transform. However, DFT is an orthogonal transformation mode, which requires huge operation, which also causes some difficulties in practical work. In order to improve this method, what is introduced and DFT is mentioned to reduce the operation workload. The most important concept of this concept is the birth of DCT and DFT. These two calculation modes can calculate faster and ensure accuracy. Especially the DCT method, which is very similar to the tobbelize matrix and closely related to human language. Therefore, DCT is the best transformation method in many people's cognition.

Because the image itself is two-dimensional, it is also necessary to use two-dimensional algorithm in the selection of specific algorithm. In specific application, two-dimensional DCT is used for image processing.

1) Definition of discrete cosine transform (DCT)

Set the size of the graph as $f(x,y)(x=0,1,2,...,M-1, y=0,1,2,...,N-1)$ to $M\times N$, Its two-dimensional DCT transformation formula can be shown in formula (1):

$$F(u,v)=C(u)C(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\cos[\frac{(2x+1)u\pi}{2M}]\cos[\frac{(2y+1)v\pi}{2N}] \quad (1)$$

Similarly, the definition of to transform (IDCT) is shown in equation (2):

$$f(x,y)=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}C(u)C(v)F(u,v)\cos[\frac{(2x+1)u\pi}{2M}]\cos[\frac{(2y+1)v\pi}{2N}] \quad (2)$$

Among：

$$C(u)=\begin{cases}\sqrt{1/M} & ,\mu=0\\ \sqrt{2/M} & ,\mu=1,2,\cdots,M\text{-}1\end{cases} \quad (3)$$

$$C(v)=\begin{cases}\sqrt{1/N} & ,v=0\\ \sqrt{2/N} & ,v=1,2,\cdots,M\text{-}1\end{cases} \quad (4)$$

Before embedding, the current digital watermark will be encrypted once. The current encryption processing is mainly to scramble the digital watermark, that is, to make the watermark image "beyond recognition" and lose its original appearance, and then embed the scrambled watermark into the carrier image, so as to further strengthen the concealment performance of the watermark, When you need to extract the watermark or view the original appearance of the watermark information, you can reverse the original scrambling rules to restore the original appearance of the watermark. If the watermark is extracted by the attacker, it is difficult to crack the original information of the watermark because he does not know the scrambling rules or the key set during scrambling, which is considered to be an error in extracting the watermark. Therefore, an excellent image scrambling technology can add more confidentiality and security performance to the watermark system. Generally, scrambling is to cha.In the two-dimensional discrete cosine transform, X and y are spatial domain values. Firstly, determine M = n, and the calculation formula can be expressed as:

$$F(u,v)=C(u)C(v)\sum_{x=0}^{N-1}\sum_{y=0}^{N-1}f(x,y)\cos[\frac{(2x+1)u\pi}{2N}]\cos[\frac{(2y+1)v\pi}{2N}] \qquad (5)$$

Inverse transformation to：

$$f(x,y)=\sum_{u=0}^{N-1}\sum_{v=0}^{N-1}C(u)C(v)F(u,v)\cos[\frac{(2x+1)u\pi}{2N}]\cos[\frac{(2y+1)v\pi}{2N}] \qquad (6)$$

Among：

$$C(u)=C(v)=\begin{cases}\sqrt{1/N} & u=v=0\\ \sqrt{2/N} & u=v=0,1,2,\cdots,N-1\end{cases} \qquad (7)$$

The direct coefficient DC is shown in formula (8):

$$F(0,0)=\frac{1}{N}\sum_{x=0}^{N}\sum_{y=0}^{N}f(x,y) \qquad (8)$$

VI. DCT DIGITAL IMAGE WATERMARKING ALGORITHM

In the meaningless watermark and meaningful watermark divided according to the content, the content conveyed by the meaningless watermark is often a sequence digital string, including pseudo-random sequence, pseudo-random binary sequence and chaotic sequence, which can be directly added to the carrier image when the watermark is embedded. However, this meaningless sequence often can not convey information, and can not express copyright information or logo, so it can not meet many needs of copyright protection. Meaningful watermark signals have various forms, which can be text, sound or image. When the creator needs to add his own copyright information or some kind of declaration to the watermark signal, he can use a meaningful watermark signal. In this way, the extracted watermark signal can also intuitively display the information. This design will study the meaningful watermark image and select the school emblem, as shown in Figure 4.

Figure 4. Watermark image

As one of the information encryption technologies, image scrambling algorithm has reversibility, and uses the adjustment and change of pixel position to produce visual confusion. After scrambling the watermark image, if you do not know the image scrambling rules, iteration cycle and other information, you can not correctly and completely obtain the original watermark image, which increases the cost and difficulty of piracy and the security and reliability of digital watermark to a certain extent. The common scrambling transformation methods in watermark images include Arnold, Baker, gray and so on. In this paper, Arnold method is mainly used to scramble the watermark image.

Arnold transform method is also called cat face method. For color images, the position of picture pixels is changed by changing the coordinates in the quadrant, and then the gray value of the picture is changed. Each change is a cat face transformation. After all the pixels of the image are transformed once by the formula, the image will be very different from the original image, which is equivalent to a new image. Then you can continue to transform the new image until the image loses its original appearance. At this time, assuming that the number of iterations is n, it is usually used as the key to extract the watermark. The cat face transformation is periodic. When the number of iterations increases, it will return to the original image. Different order images have different recovery periods to the original image.

In this simulation, the watermark pixel size is equal to 64 * 64. The value of scrambling times is 5. After encryption, the watermark signal is evenly distributed on all pixels. [4]When there is a small part of the embedded watermark image, we can see that the recovered watermark has a relatively complete structure. See the figure below for details.

Figure 5. Watermark image

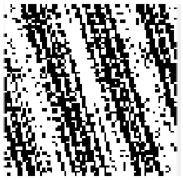


Figure 6. Scrambling once

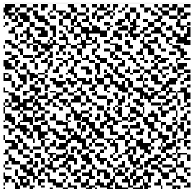


Figure 7. Scrambling 5 times

VII. Experimental Results and Anti Attack Experiment

A. Image quality evaluation

The quality of digital watermarking algorithm needs a reasonable evaluation method and index. At present, the evaluation system of digital watermarking algorithm is not perfect, and there is no unified objective evaluation standard.

Generally speaking, people can use the visual effect of the processed image to subjectively evaluate the algorithm. When the effect of the algorithm needs to be objectively evaluated and quantitatively analyzed, it needs to be completed with the help of some quantitative indicators, such as peak signal-to-noise ratio, correlation quality coefficient and normalization coefficient. A curve is a FASS curve that fills a unit square. When the image data is two-dimensional data, traversing all the elements of the image according to the trend of Hilbert curve can change the element order of the original image, so as to make the image "chaotic". Its principle is shown in the figure.

B. Analysis of experimental results

In order to verify the simulation results of the watermark, Figure 8 shows the watermark added and recovered in the carrier.[5] In the meaningless watermark and meaningful watermark divided according to the content, the content conveyed by the meaningless watermark is often a sequence digital string, including pseudo-random sequence, pseudo-random binary sequence and chaotic sequence, which can be directly added to the carrier image when the watermark is embedded. However, this meaningless sequence often can not convey information, and can not express copyright information or logo, so it can not meet many needs of copyright protection. Meaningful watermark signals have various forms, which can be text, sound or image. When the creator needs to add his own copyright information or some kind of declaration to the watermark signal, he can use a meaningful watermark signal. In this way, the extracted watermark signal can also intuitively display the information. Set the carrier image as matrix B (n, m), correspond B to each element of a one-to-one, then move the pixel at the position of element 1 in a to the position of element 2, then move its pixel horizontally to the position of element 3, translate in turn, and finally move horizontally to the position of element 1.

From the basic principle of magic square matrix, it is not difficult to see that it has periodicity like Arnold transform, and the period is. The difficulty of magic square scrambling is how to find the magic square matching the size of the image to be scrambled. Another difficulty is that when the size of the carrier image is large, the cycle of magic square transformation will be large, and the steps required to restore the image will increase, resulting in a long time-consuming algorithm.

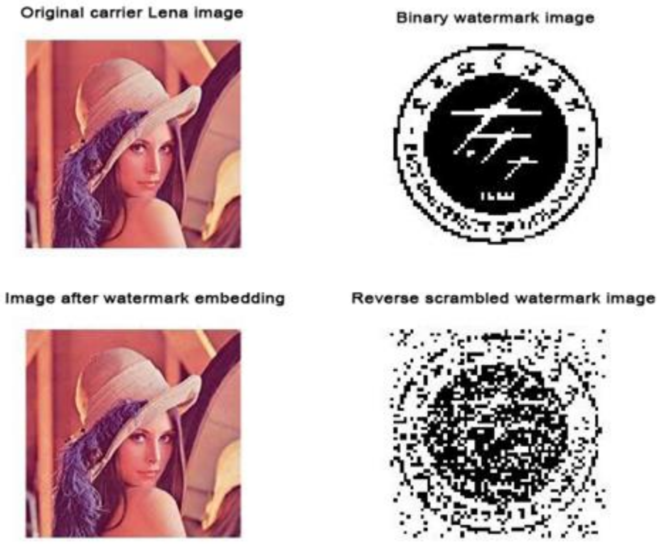

Original carrier Lena image
Binary watermark image
Image after watermark embedding
Reverse scrambled watermark image

Figure 8. Four pictures of watermark system

C. Watermark attack experiment

The robustness algorithm is mainly used to attack the watermark, and different attack methods are used to test it. The specific experimental results are shown in the figure below. In the meaningless watermark and meaningful watermark divided according to the content, the content conveyed by the meaningless watermark is often a sequence digital string, including pseudo-random sequence, pseudo-random binary sequence and chaotic sequence, which can be directly added to the carrier image when the watermark is embedded. However, this meaningless sequence often can not convey information, and can not express copyright information or logo, so it can not meet many needs of copyright protection. Meaningful watermark signals have various forms, which can be text, sound or image. When the creator needs to add his own copyright information or some kind of declaration to the watermark signal, he can use a meaningful watermark signal. In this way, the extracted watermark signal can also intuitively display the information. A point on an existing square, when changing the point (x,y) , when changing the point (x,y) to another point (x_1,y_1) has following relationship：

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (\text{mod}\,1)$$

This transformation method is also called two-dimensional Arnold transformation. The main method is to change the pixel coordinates to change the overall layout of the image. Taking the matrix distribution as an example, the application of this transformation method will mess up all the numbers in the original matrix, but if the transformation continues, there will be a chance to reset the numbers. Therefore, this method is a periodic transformation method, which is specifically expressed as:

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} (\text{mod}\,N)$$

Where is determined as a coordinate, and N is the image order. Among them, Arnold transform is used to analyze different matrices, which has a typical periodic law. In order to save cost, the content with shorter period should be selected. Magic cube originated very early, and the central idea of its transformation lies in a table lookup thinking. First, set a matrix in which the sum of two numbers on each line is equal Take natural number 1 to n ^{2}N-order matrix of elements .

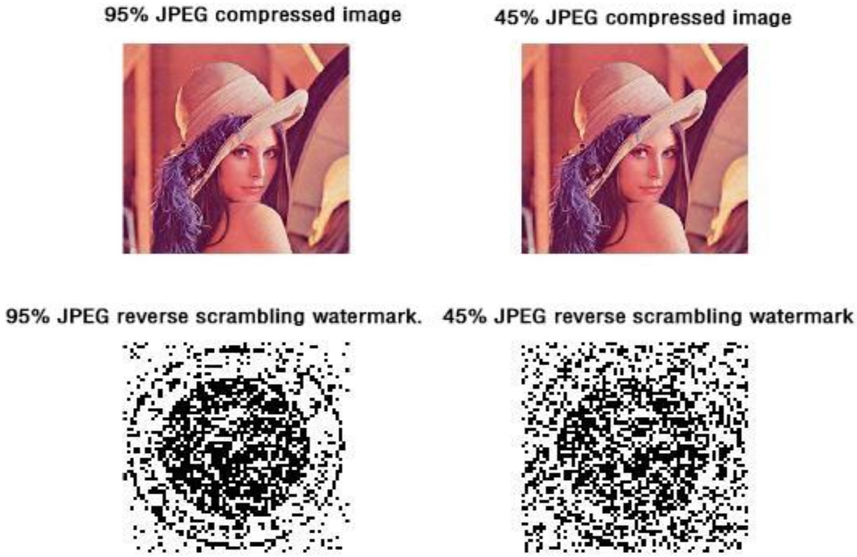

95% JPEG compressed image
45% JPEG compressed image
95% JPEG reverse scrambling watermark.
45% JPEG reverse scrambling watermark

Figure 9. Two compression test results

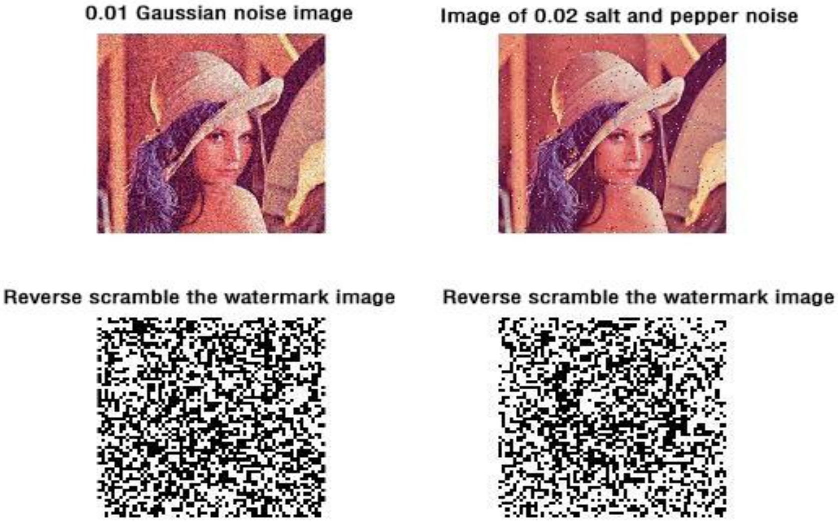

0.01 Gaussian noise image
Image of 0.02 salt and pepper noise
Reverse scramble the watermark image
Reverse scramble the watermark image

Figure 10. Noise test results

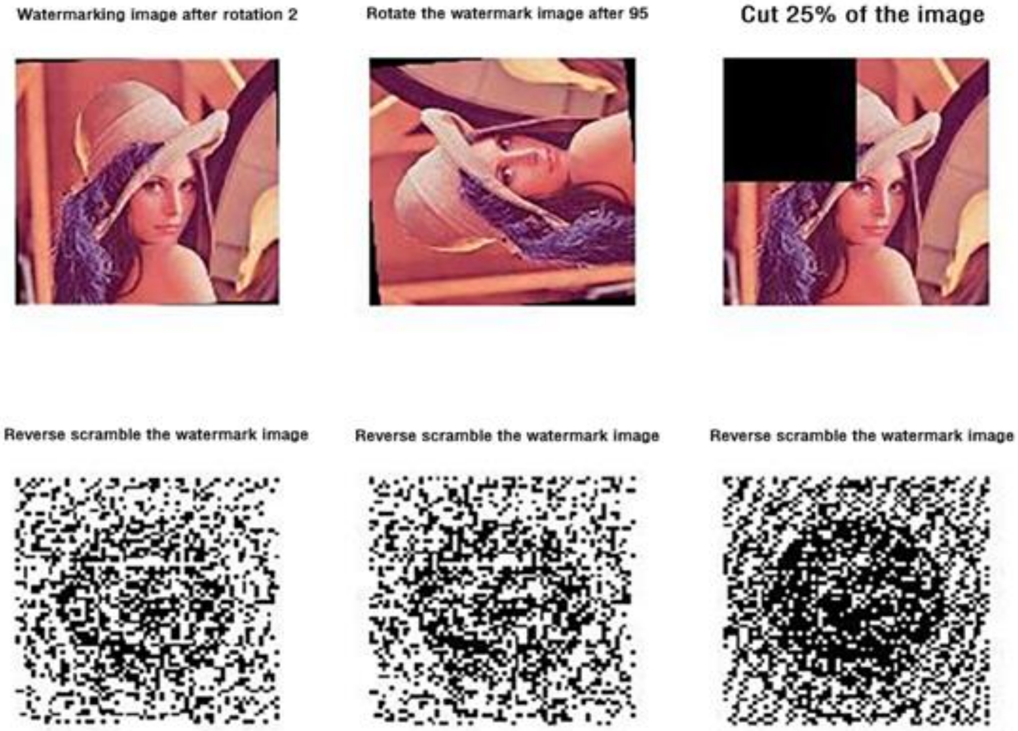

Watermarking image after rotation 2
Rotate the watermark image after 95
Cut 25% of the image
Reverse scramble the watermark image
Reverse scramble the watermark image
Reverse scramble the watermark image

Figure 11. Rotation and shear test

In the meaningless watermark and meaningful watermark divided according to the content, the content conveyed by the meaningless watermark is often a sequence digital string, including pseudo-random sequence, pseudo-random binary sequence and chaotic sequence, which can be directly added to the carrier image when the watermark is embedded. However, this meaningless sequence often can not convey information, and can not express copyright information or logo, so it can not meet many needs of copyright protection. Meaningful watermark signals have various forms, which can be text, sound or image. When the creator needs to add his own copyright information or some kind of declaration to the watermark signal, he can use a meaningful watermark signal. In this way, the extracted watermark signal can also intuitively display the information.

The following table 1 shows the relevant parameters in the attack image watermarking experiment:

TABLE I. WATERMARK SIMULATION RELATED PARAMETERS

Attack mode	Parameter	Normalized parameters	Signal to noise ratio (db)
unassailed	nothing	0.894375	38.424410
95%JPEG compress	95%	0.894085	38.293235
45%JPEG compress	45%	0.754401	35.026503
Gaussian noise	1/4	0.547445	20.157188
Salt and pepper noise	0.002	0.664663	21.999809
Rotate	2°	0.796050	23.123601
Rotate	95°	0.779734	19.440330
Shear	1/4	0.664663	11.225899

It can be seen from the above pictures that different processing methods have different effects on the restoration process of watermark. However, the correlation coefficient of the watermark recovered after the denoising algorithm is not high, and the visual effect is not good. Therefore, it can be concluded that the robustness of the simulation algorithm is not good, and the watermark extraction is good without attack. Once attacked, the recovered watermark image is not ideal.

VIII. CONCLUSION

Copyright protection of digital media is an important means to encourage creators to create continuously. With the rapid progress of digital technology, all kinds of digital resources suffer more and more infringement. Facing the frequent occurrence of infringement, more new technologies and new methods are needed to deal with and solve the problem of piracy. In this design, the digital watermarking algorithm in frequency domain is studied. The main work and results are as follows:

1) The basic theory of digital watermarking is explained, and the research status of this technology at home and abroad is summarized.

2) The research of digital watermarking algorithm in DCT domain is realized. The experimental results show that the digital algorithm is feasible in watermarking.

3) Through the experiments of non attack and robustness of digital watermark, the feasibility of DW domain digital watermark algorithm in watermark processing technology is verified.

Although this design makes a preliminary exploration and attempt on the common frequency domain image watermarking algorithms, there are other methods not involved, such as the method based on DWT. In addition, in the implementation of the watermark algorithm, the carrier image used in this paper is gray image, but in practical application, the carrier image often exists in color formats such as JPG, which needs to further improve the algorithm to adapt to more carrier image formats and further expand the scope of application.

ACKNOWLEDGMENT

This paper is about school level scientific research and innovation team construction project of East University of Heilongjiang in 2021, "Computer vision and new optoelectronic devices research team", project number HDFKYTD, 202105.

REFERENCE

[1] Zhang Yafeng, He Dandan, Li Ning. Research on digital watermarking technology based on DCT algorithm [J]. Precision manufacturing and automation, 2018, 25(04): 14-16.

[2] Sun Hanqing, Li Xiyan, Wang Guizhi, Lian Weimin. New research on watermark scrambling in dwt-dct-svd domain [J]. Laser magazine, 2019, 9(02): 110-113.

[3] Hung-Jui Ko, Cheng-Ta Huang,Gwoboa Horng,Shiuh-Jeng WANG. Robust and blind image watermarking in DCT domain using inter-block coefficient correlation [J]. Information Sciences, 2019, 15(08): 110-128.

[4] Yifeng Zhang, Yingying Li Yibo Sun. Digital Watermarking Based on Joint DWT–DCT and OMP Reconstruction [J]. Circuits, Systems, and Signal Processing, 2019, 26(04): 36-47.

[5] Mahendra M. Dixit, C. Vijaya. Image Quality Improvements Using Quantization Matrices of Standard Digital Cameras in DCT Based Compressor [J]. Journal of The Institution of Engineers (India): Series B, 2019, 35(11): 100-139.

[6] Li Yingying, Zhang Yifeng, Cheng Xu, sun Yibo. Robust watermarking algorithm based on DWT optimal multi subgraph and sift geometric correction [J]. Computer application research, 2019, 14 (06): 18-23.

[7] Gao Yejun, Wang Bing. Application of digital watermarking algorithm based on DCT transform in Military Communication [J]. Digital communication world, 2019, 8 (07): 181-190.

[8] Liang Xin. Research on color image digital watermarking algorithm based on DWT and SVD [J]. Computer and digital engineering, 2019, 3 (08): 2014-2017.

[9] Hu Ping. Research and implementation of Android based hidden digital watermarking technology [D]. Beijing University of Posts and telecommunications, 2018, 19 (41): 22-43

[10] Yu Shuaizhen. Xie Daoping. Dct-svd joint digital watermarking algorithm based on aronld scrambling [J]. Journal of Mudanjiang University, 2019, 12 (10): 116-121.

[11] Li Lei. Digital watermarking technology based on DCT transform and SVD transform [J]. Computer knowledge and technology, 2019, 23 (30): 197-199.

[12] Liu di. A digital watermarking algorithm based on discrete cosine transform and its implementation [J]. Science, technology and economy guide, 2017,25 (35): 4-5.

Communication Architecture Design and Case Study of Embedded Partition Real-Time Operating System

Penghui Ren
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: rph_0290@163.com

***Abstract*—With the continuous development of integrated modular avionics system, a large number of applications have higher and higher requirements for the operating system. However, the kernel and application process of traditional embedded real-time operating system often run at the same privilege level. A wrong operation may cause the normal operation of the whole kernel or other processes, resulting in system crash, Embedded real-time partition operating system is widely used because of its good protection ability of time partition and space partition. Because partitions are isolated, in order to carry out data communication between partitions, it is necessary to adopt the way of inter partition communication for information transmission. This paper introduces the architecture of partitioned embedded operating system, discusses the communication principle and design process between partitioned modules, and focuses on the communication mechanism of sampling port and queue port. In addition, a communication mechanism based on virtual port is used to solve the problem that the port bound by the application process in a partition cannot communicate with the communication equipment between the partition module and other partition application processes. Finally, the design process of socket communication in partitioned operating system and the sending and receiving process of data under partitioned operating system are proposed.**

Keywords-Partition Operating System; Communication Between Modules; Virtual Port; Socket Communication

I. INTRODUCTION

With the rapid development of science and technology in China and the miniaturization and specialization of embedded operating system, embedded operating system is developing from a relatively single weak function to a more professional strong function. Embedded real-time operation system (RTOS) is the core software of airborne equipment. It is widely used by foreign enterprises and companies because of its small kernel, high stability, strong real-time and tailorability. However, in these generally applicable embedded operating systems, such as VxWorks, wince, deltaos, etc., the application process and the kernel are in the same operating system at the same time. Therefore, the wrong call of an application process may cause the wrong response of the kernel or other application processes, resulting in the failure of the system to run normally. Therefore, in order to protect the system resources and avoid the impact between applications with different functions or security levels, it is necessary to independently develop an embedded real-time operating system with its own

DOI: 10.21307/ijanmc-2021-034

independent address space and no mutual influence in time cycle.

ARINC653 is the main operating system specification to meet the requirements of integrated avionics real-time operating system. The most important thing is to put forward the concept of partition [1]. Partition refers to the collection of two or more application processes with similar or related functions running on the same processor module. The implementation of partition mainly includes time partition and space partition. Spatial partition means that each partition in the operating system has its own independent address space. By using the storage manager to establish the mapping between the partition address and the actual physical address for each partition, each partition has its own independent and unique storage address to ensure that all partitions in the space are independent of each other. Time partition means that each independent partition is scheduled in rotation according to a specific cycle. The priority of each partition is the same. The operating system provides a fixed time length, which can be divided into multiple time fragments. In this fixed time, Each partition will be allocated at least one time fragment, and the partition can only be accessed within the allocated time fragment, so as to ensure the independence and correctness of application processes in each independent partition[7].

The application process of partitioned operating system is isolated from each other in time and space, so the communication between partitioned modules has become the main way of data exchange between partitions. Section 1 introduces the architecture of partitioned embedded operating system; Section 2 describes the related contents of inter partition communication, including the concept of inter partition communication, the communication mechanism of sampling port and queue port; In Section 3, a simple design of port communication between partition modules is carried out; Section 4 mainly discusses and designs the working principle, interface function and data sending and receiving process of TCP socket under partitioned operating system; Section 5 summarizes.

II. Partition Embedded Real-Time Operating System

A. System architecture

The software structure of partitioned operating system is usually divided into three layers, including application layer, operating system layer and hardware module support layer.

The application layer is a partition application developed by users and runs on the operating system.

The operating system layer mainly implements hardware independent functional services, including the basic core functions of the operating system and various configurable components to meet the needs of specific applications. The operating system also includes partition operating system and core operating system. The partition operating system is the manager of resources in the partition to realize process management, scheduling and resource allocation in the partition. The core operating system mainly realizes partition management, scheduling, inter partition communication, system fault monitoring, resource management and equipment management in the system.

The basic core of partition embedded real-time operating system provides general control services of real-time operating system, including task management, inter partition communication, interrupt / exception management, clock / timer management, cache management, user expansion, error handling, health monitoring, storage management, device management, virtual file system management and other functions.

Configurable components are components that provide specific functional requirements for different airborne software, mainly including C runtime library, VxWorks compatible interface, bit management and file system module.

The module support layer is composed of specific hardware module support software developed according to specific interface specifications, which mainly realizes the isolation between hardware and operating system layer. The module support layer mainly includes structure support package, board level support package, MSL layer debugging agent and image management.

The API for the interaction between the operating system layer and the module support layer interface is agreed by both of them. The module support layer supports the hardware support services required by the operating system layer.

The architecture of partitioned embedded real-time operating system is shown in Figure 1[2].

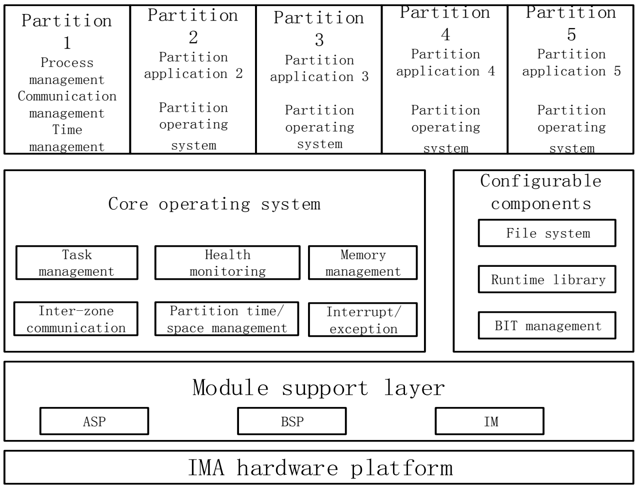


Figure 1. Partition embedded real-time operating system architecture

B. Process management

Process management is mainly responsible for the creation, scheduling and deletion of all processes in the partition. There can be two types of processes in the partition at the same time, namely, periodic processes executed at a fixed frequency and aperiodic processes triggered by events. The process states include ready state, running state and waiting state. The basic state and its changes are shown in Figure 2.

Any process can be preempted by other processes in this partition at any time. When the partition is activated, the process in the ready state is executed. Processes can lock some programs through preemption control mechanism, that is, CPU resources will not be preempted by other processes in the partition until they are unlocked. If a protected locked process in the partition is interrupted due to the end of the partition time window, ensure that the process is executed when the partition time window arrives again.

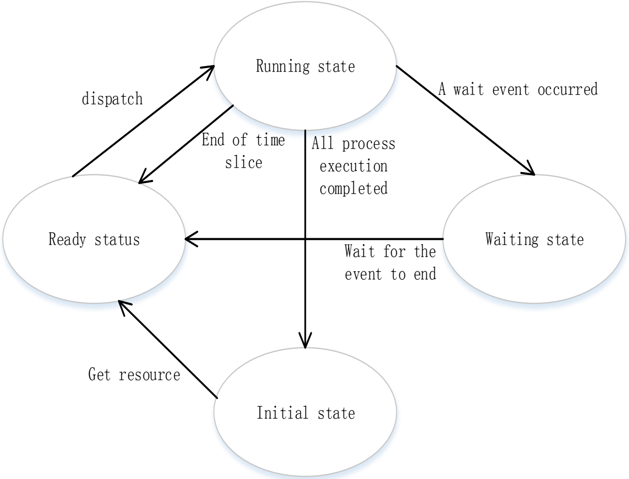


Figure 2. Process basic state and its transition diagram

C. Design objectives

The partition architecture design of separated kernel is to design the partition operating system from three aspects: partition isolation, reducing coupling and adding an intermediate layer.

1) Partition isolation

Considering the reliability design of separated kernel, the basic unit of embedded real-time operating system is task, and the resources occupied by a task are memory space and CPU time. Therefore, the kernel can be isolated from these two aspects, and different tasks can be placed in the partitions that have been isolated in time and space, so that they do not affect each other. Because each task runs in its own different partition, it does not interfere with other tasks, so as to enhance the reliability of the system.

2) Add intermediate layer

David Wheeler, a famous British computer scientist, once said the famous saying "All problems in computer science can be solved by another level of indirection." it means that all problems in the computer field can be solved by adding an indirect middle layer. The idea of adding an intermediate layer is to consider the architecture of embedded real-time operating system and provide the reliability of RTOS by adding an intermediate layer.

3) Reduce coupling

Low coupling is an important design pattern idea. On the one hand, low coupling reduces the range of other modules affected by the change of one module; On the other hand, it makes the module more cohesive, simpler structure, stronger tailor ability and easy to understand. The idea of low coupling in the design of zoning system will help to improve the reliability of the system. Taking VxWorks as an example, its wind kernel contains functions such as task management, synchronous communication and memory management. These functions are often cohesive with the device driver. The error of any module will lead to system crash. The reason is that the coupling between RTOS kernel function modules is too high, which leads to the reduction of its reliability. Therefore, the idea of low coupling is adopted to split each functional module and reorganize the RTOS structure to make it independent, so as to enhance its reliability.

In short, the core idea of partition system design is "isolation". This idea is reflected in the partition strategy of space-time isolation, the middle layer of implementation isolation, or the function block isolation to reduce coupling.

III. INTER MODULE COMMUNICATION

Communication between partition modules [8], that is, data exchange between partition modules. The only way of communication is through messages, ports and channels. The communication between partition modules is completed by sending and receiving messages through ports. Messages are sent from one source port to one or more destination ports. The port is visible to the user. Channel provides the interconnection mechanism between ports. Each channel indicates the port name and partition of sending messages, as well as the port and partition of receiving messages. The relationship between port and channel is mapped through XML configuration file. When using ports in the module, first call the port creation service to complete the creation of port objects in the partition module and realize the connection with core communication resources. The partitions communicating with each other can be in the same processor module or in different processor modules. The channel defines the logical relationship between a source port and one or more destination ports, and also defines the transmission mode and characteristics of messages from the source port to the destination port.

The communication service function between partition modules provided by the partition operating system that complies with the ARINC653 standard, on the basis of meeting the communication function between partition modules on the same module, also needs to provide support for the ability to communicate between partitions on different modules [3].

There are two types of communication services between partition modules: one is sampling mode, and the other is queue mode. The sampling mode is suitable for transmitting data messages that are generally similar and constantly updated. There is only one valid message buffer in the system, and the message remains in the buffer until it is overwritten by a newly sent message. Each module of the zone can send messages to the sampling source port at any time, or access the destination port information at any time. The queue mode is suitable for transmission that contains different data information, and does not allow the message to be overwritten, and the message is generally not allowed to be lost. The message remains in the source port until it is sent successfully, or it remains in the destination port until it is successfully received by the application port.

A. Sampling port message communication

The message of the sampling port [5] does not provide a queuing mechanism, that is to say, the sending and receiving operations will not suspend the user process, there is one and only one effective message buffer in the system, and the newly sent message will overwrite the previously sent message. The communication process of the sampling port message is shown in Figure 3.

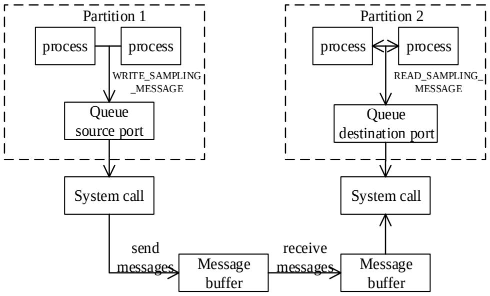


Figure 3. Sampling port message communication process

For the sender, the process calls the WRITE_SAMPLING_MESSAGE service to initiate a request to write a sampled message. At this time, the port service is checked for legitimacy, including the port ID and the legitimacy of the message. After it is legal, the user sends data to the port. If the port has no data at this time, the data is copied to the port's buffer for data transmission; if there is data in the port at this time, the original data is overwritten.

For the receiving end, the process needs to call the READ_SAMPLING_MESSAGE service to initiate a request to read the sampled message. At this time, the legality of the service (port ID and other parameters) of the destination port needs to be checked. After it is legal, the user starts to receive data from the receiving port. If there is no message on the port at this time, the message is copied to the receiving port buffer, and then the current port is empty; if there is a new message at this time, the original old message Cover, and then calculate the age of the message based on the current time and the time when the message arrives at the port, determine whether the message is valid, and finally return the validity of the message to the user.

B. Queue port message communication

The message of the queue port[5] supports a queue waiting mechanism. The operating system will create a limited message queue depth according to the situation of the message queue, maintain the state of the source port and the destination port, and then determine the processing operations that need to be performed according to the state of each port. The communication process of the queue port message is shown in Figure 4.

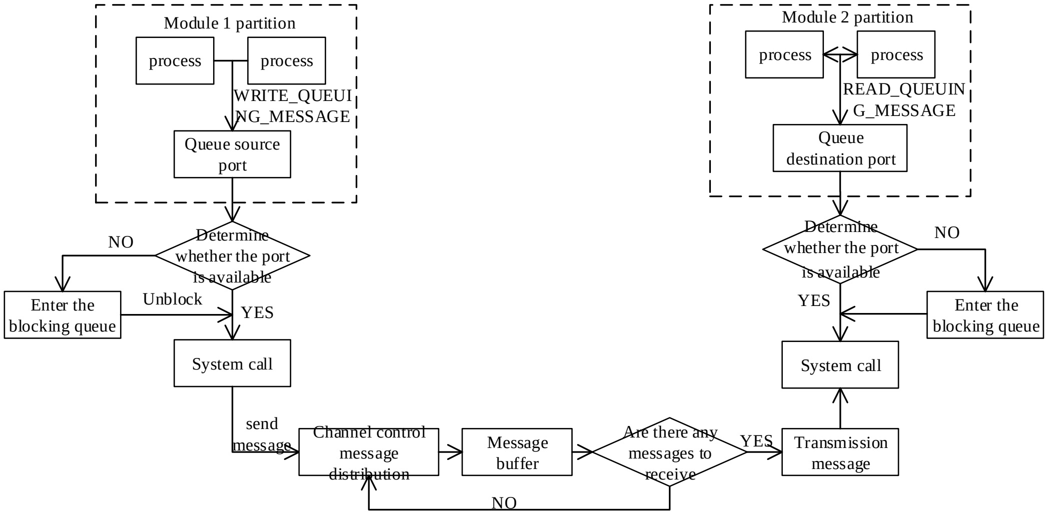


Figure 4. Queue port message communication process

For the sender, the user process calls the WRITE_QUEUING_MESSAGE service to initiate a request to send a queue message. At this time, it starts to check the legitimacy of the port, and then determines the processing strategy according to the state of the source port. If the sending port does not have a free buffer, the port is in an unavailable state at this time, and the process will enter the blocking queue; if the sending port has a free buffer, that is, the current port is in an available state, the system calls the channel control program to carry out the message distribution. The message to be sent is copied to the message buffer, and the channel control program completes the message sending operation according to the mutual connection between the source port and the destination port.

For the receiving end, the user process calls the READ_QUEUING_MESSAGE service to receive messages from the receiving port, and at the same time checks the legitimacy of the port, and then determines the subsequent operation according to the state of the destination port. If there is no message at the destination port at this time, the process will enter the blocking queue of the receiving message port; if there is a message, it proves that the current port is available, and the message is copied from the message buffer to the message queue of the corresponding destination port through a system call. Until all the destination ports have received the message, the system buffer is notified to release the space of the corresponding message buffer and the message queue of the source port.

IV. PORT COMMUNICATION DESIGN BETWEEN PARTITION MODULES

A. Port communication between partitions

In the partitioned operating system, the ports have the following types: the first is a local port, which allows the application process to communicate with other application processes on the same module, and the local port is attached to the partition of the module; the second is Virtual port, through the virtual port can communicate with the partition outside the partition module, by connecting a port to the underlying driver; the third type is direct access to the port, this port implements a queue port without a software buffer, and it can also It is directly used to communicate with the partition outside the module. However, a channel with a direct access port must have a source address and a destination address. The direct access port can be in a partition or other partitions.

The communication between partition modules, as the name suggests, is data exchange between different partition modules, which determines that the sending partition and the receiving partition will not be in the same operating system, so the concept of virtual port [4] is introduced.

Corresponding to the real port we are talking about, we can regard the virtual port as a temporary port in this module of the external module that communicates with the real port, but ultimately the message transmission between modules is completed through the underlying driver. The communication link between the partition modules is shown in Figure 5.

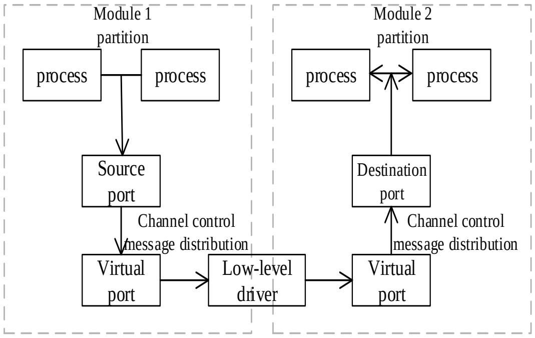


Figure 5. Communication between partition modules

For the sender, the destination port is outside the module, so we configure a virtual port for the destination port, which corresponds to the external device. After the user process sends data to the destination port, the operating system calls the underlying driver through the virtual port to send the data.

For the receiving end, the source port is outside the module, so we configure a virtual port for the source port. The virtual port calls the underlying driver to receive data, and then passes the received data to the user process through the destination port.

B. Virtual port

1) Virtual port function

The virtual port is a logical structure that connects the partition application port and the lower-layer network drive device. It connects the upper-layer source port and destination port through a channel. The function prototype is as follows.

```
STATUS portVirtualDrvAdd
(PORT_DRV_FCT *pPortDrvFct,
unsigned char *name)
```

Name is the name of the connected underlying drive device. The PORT_DRV_FCT structure contains 6 function pointers for virtual port creation (createRtn), virtual port read data (readRtn), virtual port write data (writeRtn), virtual port status acquisition (statusRtn), virtual port for port attachment (attachRtn) and virtual port validity check (availableRtn).

During the initialization of the inter-area communication resources, for the virtual port, the virtual port object is first attached to the virtual port driver by calling attachRtn. The input parameter of the attachRtn function is the related information of the virtual port object, and the output parameter is the ID assigned to the virtual port object by the attached virtual port driver. Pass this ID into the operating system as an input parameter when calling createRtn, readRtn, writeRtn, statusRtn, and availableRtn functions.

2) The logical structure of the virtual port

For the sending end, the virtual port logical structure table is used to record the information of all sending ports in the zone and provide information about the sending port of the zone.

PortID is the ID of the sending port; MsgName is the topic information of the current port; MsgMaxSize represents the maximum buffer message size of the sending port Length; MsgMaxNum represents the maximum number of buffered messages; MsgQueueID is the message queue ID of the buffered message; DestMsgQueueID represents the ID of the buffer message queue of the receiving port bound to the sending port. There can be multiple message queue IDs. When DestMsgQueueID is 0 When represents that the sending port is not bound to any receiving port; PsudoID represents the user configuration number of the current virtual port; DevHdr represents the device handle of the underlying driver port; EmptyFlag is a flag for judging whether the current virtual port message buffer is empty. The core operating system sends the data of the message queue of the sending port buffer to the message queue of the receiving port according to the information of the sending port and the virtual port.

For the receiving end, the logical structure table of the virtual port is used to record all the receiving port information in the partition and provide the information of the receiving port in the partition. PortID is the ID of the receiving port; MsgName is the subject information of the receiving port; MsgMaxSize represents the maximum length of messages buffered by the receiving port; MsgMaxNum represents the maximum number of buffered messages; MsgQueueID is the message queue ID of the receiving port buffer; IsAssign represents whether the port is Bind to the sending port, 1 means binding, 0 means unbound. The core operating system binds the virtual port and the port according to the information of the sending port and the information of the receiving port to establish a data channel for communication. The partition application receives data from the port buffer based on the PortID.

The logical structure of the virtual port is shown in Table 1.

TABLE I. LOGICAL STRUCTURE OF VIRTUAL PORT

content	Function
PortID	Port ID (receive/send, the same below)
MsgName	Subject information
MsgMaxSize	Maximum length of buffered message
MsgMaxNum	Maximum number of buffered messages
MsgQueueID	Message queue ID
DestMsgQueueID	The ID of the buffer message queue of the receiving port bound to the sending port (sending)
PsudoID	Virtual port user configuration number
DevHdr	Device handle of the underlying driver port
EmptyFlag	Whether the message buffer is empty
IsAssig	Whether to bind with the sending port (receive)

V. COMMUNICATION EXAMPLE

A. Socket overview[6]

Socket provides three types of sockets, namely streaming sockets, datagram sockets and raw sockets. Streaming sockets provide a connection-oriented, reliable data transmission service. The data is sent without errors and repetitions, and is received in the sending order, using the TCP protocol. The datagram socket provides a connectionless service. The data packet is sent in the form of an independent packet without error-free guarantee. The data may be lost or duplicated, and the receiving sequence is disordered. The UDP protocol is used. Raw sockets are often used to test the implementation of new protocols or access new devices configured in existing services. This interface allows direct access to lower-level protocols such as IP and ICMP. This article mainly introduces the streaming socket mode.

B. Socket layer design

In each partition operating system, the Socket layer completes the communication between data by calling the Socket interface [9] provided by the kernel TCP/IP protocol stack [11]. The specific interfaces and functions used are shown in Table 2.

TABLE II. INTERFACE FUNCTION AND CORRESPONDING FUNCTION

Interface function	**Function**
socket( )	Create a socket descriptor
bind( )	Bind the socket to a specific TCP port
listen( )	Listening socket
connect( )	Send a connection request to the server (client-only)
accept( )	Accept connection request (server exclusive)
send( )	send data
recv( )	Receive data
close( )	Close socket

1) Create socket

Function prototype:int socket(int domain,

int type,int protocol);

Create a socket to complete the following tasks:

a) Set protocol family；

b) Specify the socket type；

c) Specify the protocol related to the socket type.

2) Bind socket

Function prototype: int bind (int sockfd, const struct sockaddr *addr, socklen_t addrlen);

The main task of binding a socket is to assign a specific address (ip address + port number) in an address family to the socket.

3) Listening socket

Function prototype: int listen (int sockfd, int backlog);

The created socket is of an active type by default. The task to be completed by the monitoring socket is to change the socket to a passive type and wait for the client's connection request.

4) Send connection request

Function prototype: int connect (int sockfd, const struct sockaddr *addr, socklen_t addrlen);

The client establishes a connection with the TCP server by calling the connect function.

5) Accept connection request

Function prototype: int accept (int sosckfd, struct sockaddr *addr, socklen_t *addrlen);

After the TCP server listens to the connection request sent by the client, it will call the accept function to receive the request, thereby successfully establishing the connection. After that, the network I/O operation is started.

6) Data sending

Function prototype: ssize_t send (int sockfd, const void *buf, size_t len, int flags);

Data transmission completes the following tasks:

a) Receive the data submitted by the virtual port into the send buffer；

b) Determine the destination IP address；

c) Determine the destination port；

d) Send data。

7) Data reception

Function prototype: ssize_t recv (int sockfd, void *buf, size_t len, int flags);

Data reception completes the following tasks:

a) Receiving data on the bound TCP port;

b) Submit the received data to the virtual port first, and then submit it to the application buffer through the virtual port buffer.

8) Close socket

Function prototype: int close (int socketfd);

After the client and the server complete the data receiving and sending operations, the corresponding socket descriptor is closed, and the descriptor can no longer be used by the calling process.

After the partition operating system is started, task A of partition 1 enters the processing and waiting, and waits for the response of the network card driver and the IPC messages from other partitions in turn. The tasks of other partitions start to establish sockets to prepare for network communication. First call v_socket, v_bind and other interfaces to establish a socket connection, and then call v_listen or v_connect to monitor or establish a connection. If the application is a server, after listening to the request information Call v_accept to accept the request, send a socket command request to the task of partition 1 through SYN_SEND, call the v_recv/v_send function for network communication, and call SYN_RECV to wait for the result returned by task A. Other tasks of partition 1 receive service requests such as v_socket and v_bind from other partitions and are activated to create socket devices and establish communication, and then send and receive network data with other partitions, and then send socket handles and network addresses and other information Reply to a certain task B of other partitions through SYN_SEND, and the socket communication enters the ready state at this time. Task B sends and receives network information through the socket. All the processes are the same as the execution of task A. After receiving the reply from task B, a socket communication is completed.

C. Data sending and receiving process[10]

1) Data sending process

When sending data to the receiving end, the application process first obtains a socket, searches for the corresponding virtual port according to the IP address and port number bound to the socket, and then determines whether the connection is a TCP connection. If so, submit the data to be sent to the virtual port, and then call the send() function to send the data on the virtual port to the virtual port of the receiving end, Finally, judge whether the transmission is successful. If successful, return a parameter value of successful transmission. The process is shown in Figure 6.

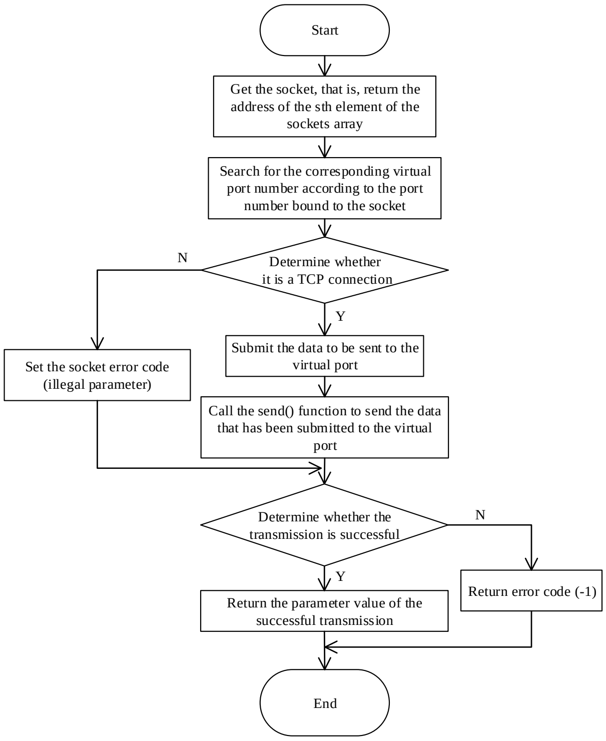

Start
Get the socket, that is, return the address of the sth element of the sockets array
Search for the corresponding virtual port number according to the port number bound to the socket
Determine whether it is a TCP connection
N
Y
Set the socket error code (illegal parameter)
Submit the data to be sent to the virtual port
Call the send() function to send the data that has been submitted to the virtual port
Determine whether the transmission is successful
N
Y
Return the parameter value of the successful transmission
Return error code (-1)
End

Figure 6. Data sending process in partitioned operating system

2) Data receiving process

When the application process receives data from the sender, it first obtains the socket, that is, selects the address of an element from the sockets array, returns its socket structure, and judges whether the connection is a TCP connection. If so, according to the socket binding Search for the corresponding virtual port number for the IP address and port number, then call the recv() function to receive data from the sender, and then determine whether the data is received, if it is received, submit the received data to the virtual port buffer, and finally the virtual port The data on the port is transferred to the port bound to the socket to complete the data reception. The process is shown in Figure 7.

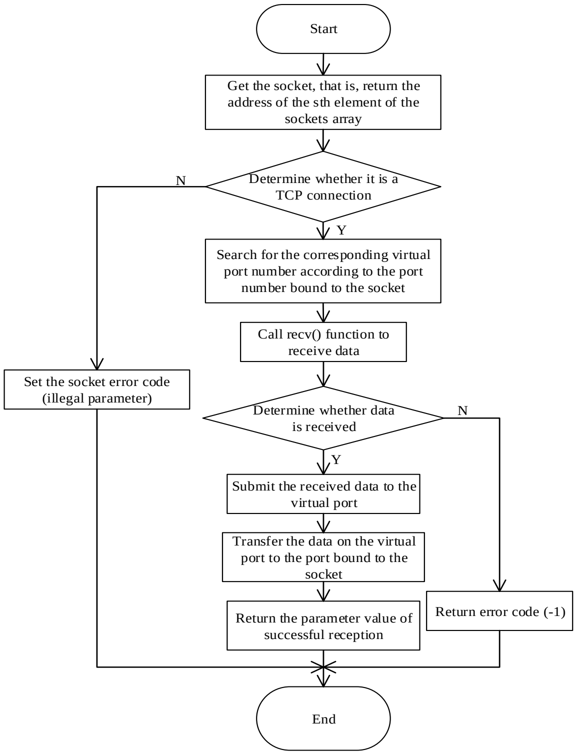

Start
Get the socket, that is, return the address of the sth element of the sockets array
Determine whether it is a TCP connection
N
Y
Search for the corresponding virtual port number according to the port number bound to the socket
Call recv() function to receive data
Set the socket error code (illegal parameter)
Determine whether data is received
N
Y
Submit the received data to the virtual port
Transfer the data on the virtual port to the port bound to the socket
Return the parameter value of successful reception
Return error code (-1)
End

Figure 7. Data receiving process in partitioned operating system

VI. SUMMARIZE

Through the analysis of the embedded real-time partition operating system architecture, this paper discusses the process management mode in the partition system and the design goals of the partition operating system, introduces the communication principle between partition modules, and focuses on the message of the sampling port and the queue port. Communication principle, and then design the port communication process between the partition modules, and realize the port communication by introducing the virtual port. At the same time, the virtual port function and logical structure are introduced in depth, and then the actual TCP socket layer of the partition operating system is introduced. The communication process is designed, using the basic function interface used by the traditional operating system, and finally a process of sending and receiving data under the partition operating system is proposed. Through an example to test whether the socket communication based on the virtual port mechanism of the partition operating system can be carried out, a simple communication result is obtained. In the future research work, we will continue to optimize the communication process of this design and the process of sending and receiving data. Although the communication between partitions has many advantages, if we do not understand and avoid the risks and problems that may be caused by the communication between partitions, it will inevitably bring many unforeseen problems in the future design work. Therefore, it is necessary to standardize and strictly design the communication process between the partition modules to reduce the risk, so that it can provide users with more powerful communication support.

REFERENCES

[1] Tao Yongchao, Song Qilong, Piao Songhao. Design and implementation of partition Operating System based on ARINC653 standard [J]. Journal of Physics: Conference Series, 2021, 1732(1).

[2] Tong Yan, Yuan Haofang, Xu Fei, Wu Zhiming, Wang Manda. Research on partition operating system based on ARINC653 [J]. Electronic Testing, 2020(13).

[3] Xu Xiaoguang, Ye Hong. Design and Implementation of Interval Communication in Avionics System [J]. Aeronautical Computing Technology, 2005.

[4] Zhang Ming, Zhou Lin.Design and implementation of IMA based on VxWorks653 partition operating system [J]. Firepower and Command Control, 2014.

[5] Xu Xiaoguang, Yun Haishun, Xing Liang. The design of inter-partition communication under partition operating system [J]. Modern Electronic Technology, 2013.

[6] Zhang Xiaona, Chang Leran, Wu Wei, Liao Jinwei, Shen Liwen. Realization of Socket Communication under Linux System [J]. Electroacoustic Technology, 2020.

[7] Huang Runlong, Shen Qian, Gou Xiantai. Task scheduling of ARINC653 multi-core and multi partition operating system [J]. Telecommunications technology, 2020, 60(09):1108-1113.

[8] Yang juping. Research on partition technology based on embedded real-time operating system [J]. Industrial control computer, 2015, 28(05):29-30.

[9] Xiao Lei. Network communication design based on socket under VxWorks [J]. Computer and network, 2013, 39 (12): 66-68

[10] Xing Liang, Zhao Yi. Application design of socket communication under partitioned operating system [J]. Aviation computing technology, 2011, 41(05):88-90.

[11] Wang Xiaopeng. Socket and Winsock communication mechanism under TCP / IP [J]. Aviation computing technology, 2004 (02):126-128.

Research on the Application of Agent-based Real-time Monitoring System in Inference Engine Cluster

Xu Jiangtao
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: 1400693814@qq.com

Yang Bo
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail:16764496@qq.com

Liu Pingping
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: 134369601@qq.com

***Abstract*—In order to ensure the stable, reliable and healthy operation of spacecraft, this paper introduces the spacecraft expert diagnosis system for real-time monitoring and early warning of spacecraft. The inference engine is the core of the spacecraft expert system. Each inference engine is responsible for a spacecraft monitoring and early warning. How to ensure the stable and reliable operation of the inference engine is an urgent problem to be solved. In this paper, agent technology is introduced to monitor the reliable operation of the inference engine and process migration when the inference engine fails. At the same time, monitor server resource utilization for process scheduling to achieve server load balancing. Practice has proved that the system can greatly improve the stability of the inference engine, improve the efficiency of spacecraft management, and save a lot of manpower and material resources.**

Keywords-Multi-reasoning Machine; Process Scheduling; Reliability; Agent; Monitoring

I INTRODUCTION

With the rapid development of China's aerospace industry, the monitoring and management of spacecraft has become more and more important. Spacecraft monitoring parameters as few as five or six hundred, as many as several thousand, each parameter anomaly may affect the normal operation of the spacecraft system, and even lead to the paralysis of the spacecraft, thus bringing immeasurable losses to the country. In the traditional management of spacecraft, the spacecraft managers monitor the operation parameters of the spacecraft, and send them to the spacecraft experts after finding the abnormal parameters. The spacecraft experts analyze the abnormal parameters, and then give the treatment scheme. Due to the small number of spacecraft, it can better complete the spacecraft monitoring and management tasks. However, with the increase in the number of spacecrafts, the mode of manual monitoring and analysis can no longer meet the needs of spacecraft monitoring and management.

So the spacecraft expert diagnosis system [2] [4] [5] platform is introduced. The knowledge of the spacecraft expert [3] is expressed as the knowledge that the inference engine can handle. The knowledge is loaded when the inference engine [1] runs, and the diagnosis results are given combined with the current state of the spacecraft operation parameters. In the specific implementation, each inference engine is responsible for real-time processing of the operating parameters of a spacecraft, so that all spacecrafts form a cluster of inference engines [6] [7]. Obviously, the stable operation of the inference engine is directly related to the safety of

DOI: 10.21307/ijanmc-2021-035

the spacecraft. So, how to make these inference engine cluster reliable and stable operation is an urgent problem to be solved, this paper gives practical and feasible solutions to this problem. Practice has proved that the system can not only greatly improve the stability of the inference engine, but also optimize the system resources and dynamically schedule the inference engine process, thus greatly improving the efficiency of spacecraft management and saving a lot of manpower and material resources.

At present, major aerospace powers are carrying out research on spacecraft fault management technology. NASA has made spacecraft failure technology the first requirement of space flight technology in the 21st century. NASA's spacecraft fault management projects include X-33 / X-34 / X-37, and military aircraft projects include F-18, F-22, JSF, UCAV, etc. The fault management technology in the United States mainly has two representatives: 1 integrated fault management technology (IVHM) for carriers, 2 prediction and fault management technology (PHM). At present, there is still a big gap between our country's research on spacecraft fault management technology and world technology, and there is no comprehensive research and verification of spacecraft fault management technology. On the one hand, our overall scientific and technological level is lower than that of developed countries. On the other hand, because of the small number of spacecrafts, the current ground measurement and control system can still be maintained. However, with the increase in the number of spacecrafts, the contradiction of low management level highlights. How to solve this contradiction and improve the management level of spacecrafts by using spacecraft fault knowledge management technology has become the only way for us.

II SYSTEM STRUCTURE

According to the characteristics of inference engine cluster real-time monitoring software running environment, this paper adopts the following system structure.

A. *Network Topology Architecture*

The spacecraft expert diagnosis system platform [8] [9] is running on the dedicated network. Including: several servers, several clients and database servers. The inference engine process runs on the server. In order to ensure the stable operation of the inference engine process, in general, a server installs a inference engine process and runs a inference engine process. The network topology is shown in Figure 1.

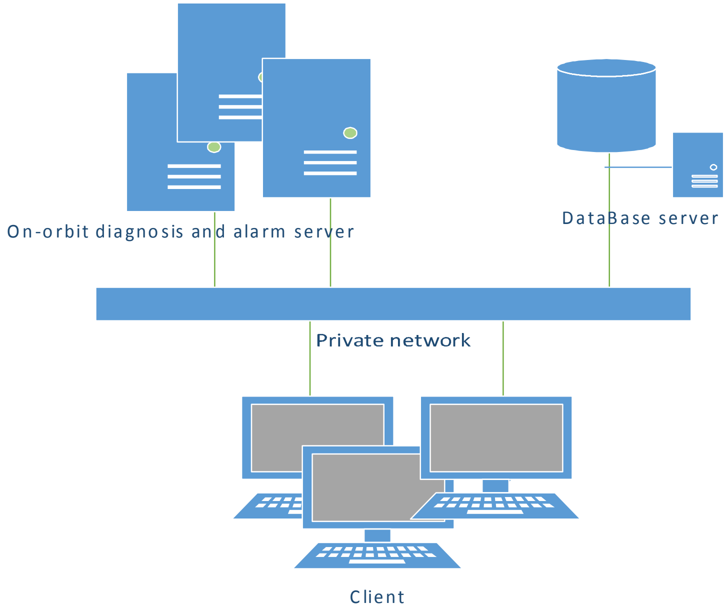


Figure 1. Topology Structure Diagram of Private Network

When multiple servers participate in the actual work, the main monitor disperses the huge data processing business to each server to achieve load balancing, so that multiple servers process the corresponding data at the same time, and thus can be competent for a large number of data processing tasks under the new demand. In each server mode in the LAN, even if the amount of data information increases again later, it only needs to be solved by task scheduling method or increasing the number of servers.

B. Factors Affecting Scheduling Structure

To realize the multi-satellite fault diagnosis task division, the scheduling structure must be formulated firstly. The scheduling structure is the basis of designing the optimization model and solving algorithm. The scheduling structure of the multi-satellite fault diagnosis system must fully consider the characteristics of the problem and the requirements of the system in terms of scalability and maintainability, including:

1) Constraint Representation of Satellite Telemetry Data and Processing

Single star fault diagnosis task planning is only for a star telemetry data analysis. When there are multiple satellites, due to different satellite data formats and different time constraints on telemetry data analysis, it must be completed within the specified time, otherwise it will lead to disastrous consequences. It is difficult to uniformly represent the constraints of multi-satellite fault diagnosis system. Only some important constraints are considered, and some secondary constraints are ignored.

2) System Scalability

With the continuous development of satellite performance and quantity, multi-satellite fault diagnosis system task planning must have good scalability to facilitate the dynamic loading and reconstruction of satellite fault diagnosis system.

3) Efficiency and Effect of Scheduling Algorithm

The task scheduling of multi-satellite fault diagnosis system has the characteristics of combinatorial optimization problem, which has been proved to be NP-hard problem. When the number of tasks and satellites increases, it is very difficult to solve the problem. Multi - star scheduling structure must consider the design of optimization algorithm, so that it can obtain satisfactory solutions with certain quality in a certain time range. In addition, the system should also consider factors such as cost and maintainability.

C. Comparison of Scheduling Structure

Aiming at the task scheduling problem of multi-satellite fault diagnosis system, this paper studies the scheduling structure of multi-satellite fault diagnosis. At present, there are mainly two scheduling structures: centralized scheduling structure and hierarchical scheduling structure.

1) Centralized Scheduling Structure

Centralized scheduling structure: There is a scheduling host responsible for collecting system load information in the system, which is the main body of centralized scheduling. It controls the solution and keeps synchronization with the information interaction between multiple computing hosts. The host maintains a task allocation table and assigns tasks according to the system load. Other hosts are computational hosts, and computational hosts are only responsible for processing telemetry data.

The advantages of this strategy are: the scheduling host has global information, easy to make decisions and maintain load balance, easy to track execution. The algorithm is easy to implement, suitable for the network environment with fewer nodes, and has better performance on bus network. The disadvantage of this strategy is that the computing host needs to wait when scheduling the host to collect information of all hosts, which wastes the processing power of the computing host.

2) Hierarchical Scheduling Structure

The core idea of the hierarchical scheduling structure is that the task pre-scheduling center allocates tasks to different scheduling modules according to the amount of satellite telemetry data. Each scheduling module handles the fault diagnosis of one satellite, and then each scheduling module allocates tasks to different servers.

The advantage of this strategy is that the multi-satellite fault diagnosis task scheduling is transformed into multiple scheduling module problems by pre-allocation of tasks, which reduces the difficulty of solving the problem. Since each scheduling module only conducts fault diagnosis task scheduling for one satellite, the problem scale is small, and the fault diagnosis task scheduling of multiple satellites can be operated in parallel, so the efficiency of problem solving is high. The downside is that additional scheduling hosts are needed to increase costs.

D. Scheduling Structure Selection

In order to meet the scalability and solution requirements of multi-satellite fault diagnosis system, this paper proposes a distributed scheduling structure. It includes core scheduling modules and extensible modules.

1) Core Scheduling Module

The subject of centralized scheduling performs solution control and information interaction with multiple extended modules and maintains synchronization.

2) Extensible Module

It is responsible for fault diagnosis of telemetry data of multiple satellites. When there are additional satellites and the load of each server is large, the number of servers can be increased to solve the problem.

E. Basic Principle of System Implementation

Real-time monitoring software inference engine [10][12][13] to achieve the basic principles shown in Figure 2.

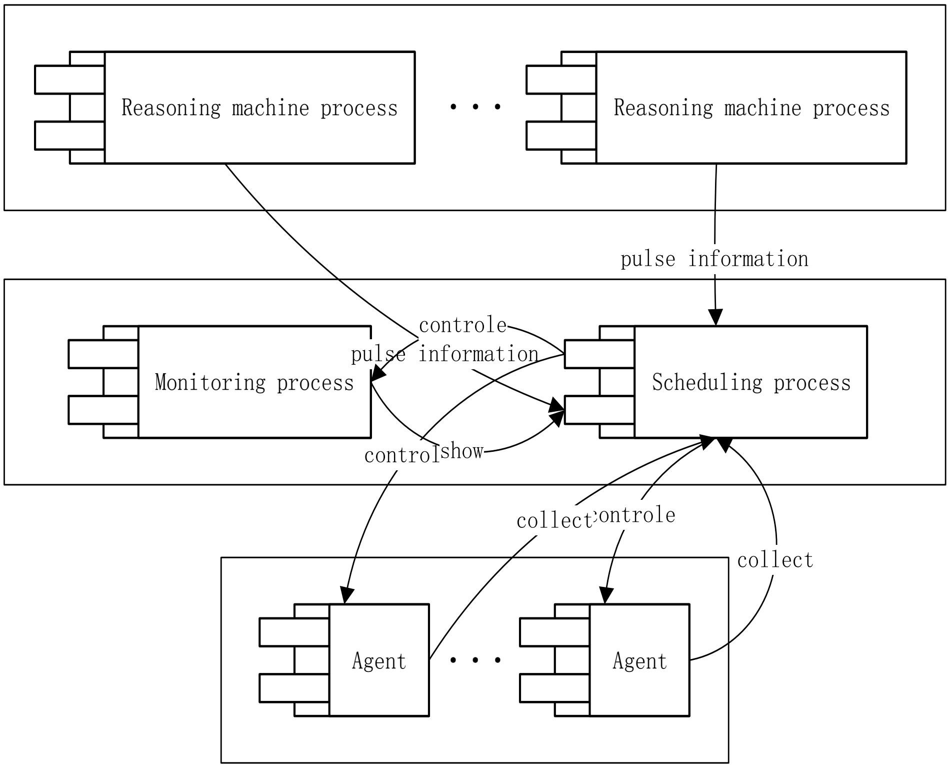


Figure 2. Basic Principle Diagram of System Implementation

The inference engine real-time monitoring software includes: system monitoring client process, process monitoring scheduling process, process monitoring service process. The basic principle of the system operation is as follows:

1) Pulse Information

Each monitored inference engine process sends its own pulse information. The pulse information is sent once every 1 second and broadcast by UDP communication. If the process monitoring process can receive the pulse information within 5 seconds, the target process is running normally; otherwise, deal with the target process running abnormally.

2) Process Monitoring Service Process

The system monitoring service process starts automatically as the operating system starts. The system monitoring service process is responsible for starting the system scheduling service process, starting the process, stopping the process, downloading and installing the software package, obtaining the system running load information, obtaining the process running status information, obtaining the system running log information, monitoring the process running status, and broadcasting the process pulse information.

3) System Monitoring Client

System running status monitoring client mainly includes: management server running information, management process running information, management exception handling and TCP communication. The management of server operation information is mainly responsible for the display and maintenance of server operation parameters; managing process running information is responsible for displaying process running information, managing process, etc. Managing exception handling mainly deals with system running exceptions; TCP communication module is used to communicate with the process of process scheduling subsystem, send operation commands and receive operation results.

4) Process Monitoring Scheduling Process

The system monitoring service process starts with the operating system, and then starts the system scheduling service process according to the 'election' algorithm. The process monitoring scheduling process is mainly responsible for monitoring the running state of the target process, monitoring the server running information, maintaining system load balancing, and managing the target process. The following details:

a) Monitoring the Running State of the Target Process

The process monitoring and scheduling process receives the pulse information of the target process in real time, and continues if the pulse information of the target process is received in timeout; otherwise, set the run state of the target process to a ' fail ' state.

b) Monitor Server Running Information

The process monitoring scheduling process can send ' Get Server Average Load ' or ' Get Server Real Time Load ' commands to the process monitoring service process of the monitored server. Send server load information when the process monitoring scheduling module requires server load information.

c) Maintenance of System Load Balancing

Process monitoring scheduler selects the target process to migrate, the source server to migrate, and the target server to migrate according to the scheduling algorithm. See Figure 3 for specific scheduling algorithms and process migration processes.

d) Management Target Process

The management target process includes: starting the target process, stopping the target process, migrating the target process, downloading the installation server software package, upgrading the server software package, process master-slave switching, obtaining the system running load information, obtaining the process running state information, obtaining the system running log information, monitoring the process running state, broadcasting the process pulse information.

Process monitoring scheduling process realizes process scheduling according to scheduling algorithm and maintains system load balancing. Process scheduling has two ways: automatic scheduling and manual scheduling. The automatic scheduling algorithm [11] is shown in Figure 3.

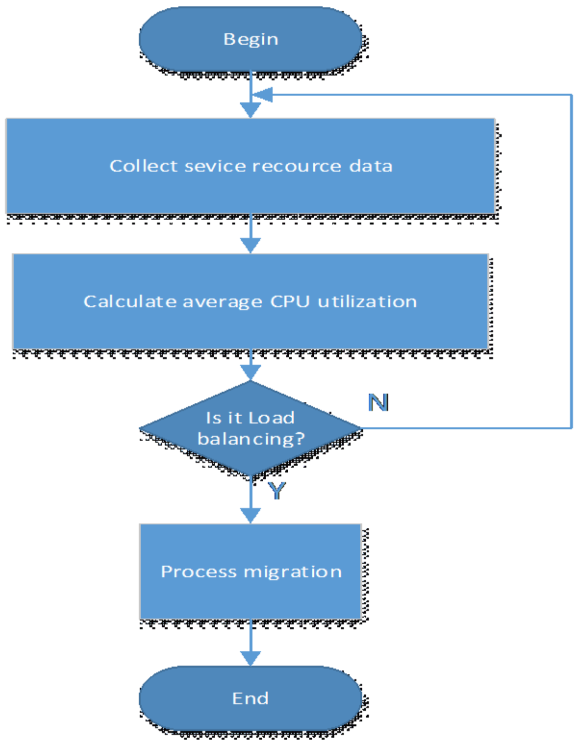

Begin
Collect sevice recource data
Calculate average CPU utilization
Is it Load balancing?
N
Y
Process migration
End

Figure 3. Process Scheduling Algorithm

The average CPU utilization formula is as follows:

$$\overline{U} = \frac{1}{N}\sum_{i=1}^{N} u_i \qquad (1)$$

Where: $\overline{U}$ represents average CPU utilization, u_i represents the ith CPU utilization, N represents the number of CPUs owned by the server.

Manual scheduling is the process and server of migration selected by spacecraft users. At this time, the migration mode of the migrated process is set as 'manual migration'. Then, the system scheduling service process migrates according to the above scheduling process.

III Software Architecture

In order to realize the running goal of all servers and service processes on the monitoring network, the system is divided into three parts: system running status monitoring client, system scheduling service process, system monitoring service process. System running status monitoring client running in one or more clients; the system scheduling service process runs on a certain server and is initiated by the system monitoring service process; the system monitor service process runs on each monitored server and starts with the operating system. The working process of inference engine real-time monitoring software is shown in Figure 4.

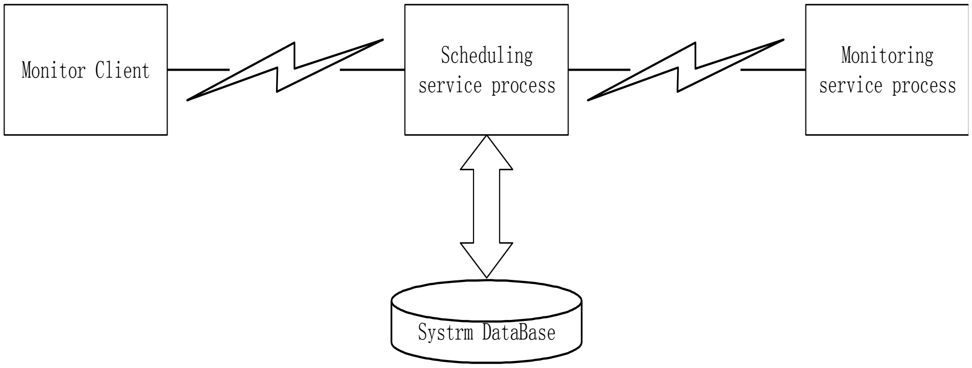


Figure 4. Software Architecture Diagram

The monitoring system status information consists of three parts: system operation status monitoring client, system scheduling service process and system monitoring service process.

A. System Running Status Monitoring Client

System running status monitoring client main modules: server management module, process management module, exception handling management module and TCP communication module. The server management module is mainly responsible for the display and maintenance of server operating parameters; process management module is responsible for displaying process running information, managing process, etc. Exception processing module mainly deals with system operation exception; TCP communication module is used to communicate with the process of process scheduling subsystem, send operation commands and receive operation results.

1) Display System Status Information

System running state information mainly includes: server running information, service process running information, etc. TCP communication module receives service process operation information, and process management module is responsible for displaying service process operation information to spacecraft users; TCP communication module receives server operation information, and server management module is responsible for displaying server operation information to spacecraft users.

2) Display System Abnormal Alarm Information

TCP communication module receives system exception alarm information, exception handling module is responsible for alarm and display exception information to spacecraft users.

3) Handling System Operation Anomalies

When spacecraft users process system exception alarm information, if they choose 'save exception handling process', exception handling module sends exception handling process information in XML form to process scheduling subsystem through TCP communication module.

B. System Scheduling Service Process

System scheduling service process mainly consists of TCP communication server module, command manager, command processor, information manager, information processor, TCP communication client module, process scheduling management module and master-slave switching module. TCP communication server module is responsible for communication with all the system running status monitoring client; command manager is used to save the received operation commands; command processor is responsible for reading and parsing commands; information manager is used to save system operation information; information processor is responsible for reading and parsing system operation information; TCP communication client module is responsible for communication with all process monitoring service modules; the process scheduling management module is responsible for starting and stopping processes.

1) Receiving Monitoring Client Operating Commands

The receiving and monitoring client operation commands are completed by independent threads. Monitors communication between client and system scheduling service processes via TCP. The monitoring client sends the monitoring client operating commands to the system scheduling service process through the TCP communication module. The TCP communication server module of the system scheduling service process is responsible for receiving the monitoring client operating commands. Monitor client operating commands received and written to the command manager.

2) Handling Monitoring Client Commands

The command processor is completed by independent threads. Command processor scan command manager. If there is an unprocessed command, read the command and parse it, and then select a TCP communication client to send the command to the system monitoring service process; if the command manager is empty, then wait for 1 millisecond to continue scanning.

3) Receiving System Monitoring Service Process Operation Information

Receiving system monitoring service process running information with independent threads to complete. After receiving the running information of the system monitoring service process, the TCP communication client module is written to the information manager.

4) Processing System Monitoring Service Process Operation Information

The information processor is completed by independent threads. Information processor scan information manager. If there is an unprocessed information, the information is read and analyzed, and then a TCP communication server is selected to send the information to the monitoring client; if the information manager is empty, then wait for 1 millisecond to continue scanning.

5) *Process Management*

The system scheduling service process has two states: the main process state and the standby process state. When the main process runs, the standby process runs in an inhibitory manner; when the main process fails, the standby process switches to the main process state. When the scheduling process is in the host state, the process of switching the standby state is shown in Figure 5.

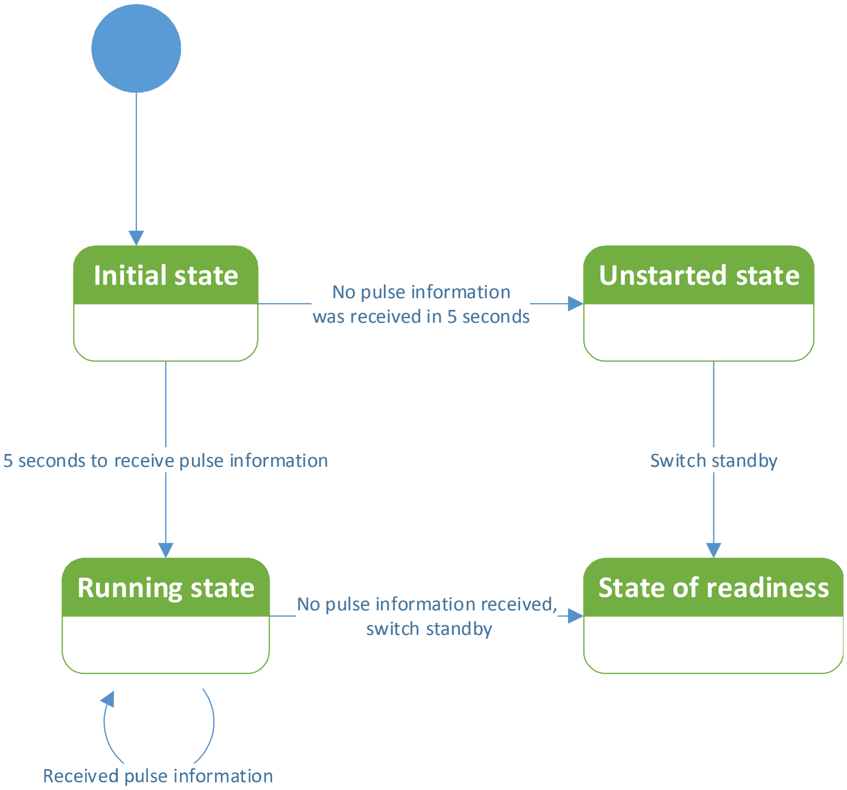

Initial state
No pulse information was received in 5 seconds
Unstarted state
5 seconds to receive pulse information
Switch standby
Running state
No pulse information received, switch standby
State of readiness
Received pulse information

Figure 5. Host Switch to Standby Process

After the system monitoring service process starts, it enters the 'standby' state. If the host information is not received for 5 seconds, it enters the 'election' state; automatic transition to 'counting' status after 10 seconds from 'election' status; in the 'counting' state, if the 'minimum load', broadcast 'I am the host', 5 seconds into the 'host' state; if the 'load is large' then into the 'standby' state; if you receive 'I am the host' in 'Host' state, go to 'Standby' state.

When the scheduling process is standby, the process of switching host state is shown in Figure 6.

When in the 'initial state', if the scheduling module receives the pulse information to the 'running' state, if the scheduling module does not receive the startup information to the 'not running' state; when in the 'run' state, if the scheduling module does not receive the pulse information to the 'abnormal' state; when 'not started', the scheduling module selects the server to go to 'ready to start' state, if the scheduling module does not receive the start information go to 'clean' state;

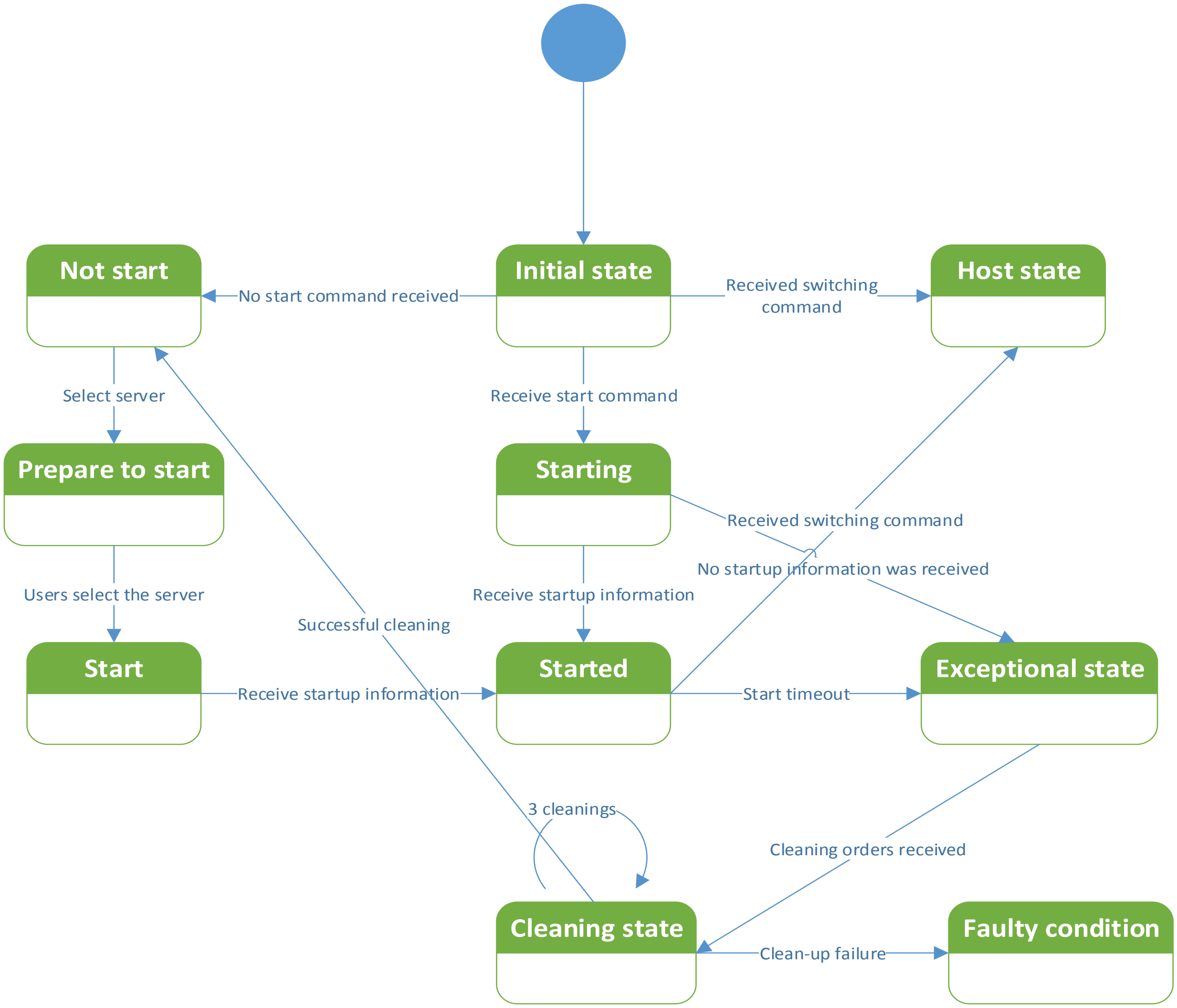


Figure 6. Standby Switch to Host Process

When 'ready to start', the scheduling module selects the server to go to the 'start to start' state. If the server is not installed, the scheduling module goes to the 'fault' state. When 'start', the scheduling module sends the start command to the 'start' state, and if the server is selected to have no software package installed, the scheduling module is transferred to the 'fault' state; when in the 'start', if the scheduling module receives the start information to the 'start' state, otherwise into the 'abnormal' state; when in the 'abnormal' state, the scheduling module directly turns to the 'clean' state; when in the 'clean up' state, if the clean up 3 times failed to go to the 'failure' state; when in the 'fail' state, when the user discovers a failure, manually set to the 'initial' state.

6) *Send Process Running Status Information*

When the monitoring client needs the process running state information, the command to obtain the process running state is sent to the system scheduling service process. The system scheduling service process command processor parses the command, and then obtains the process running state information according to the command parameters, and finally sends it to the monitoring client in XML form through the TCP communication server.

7) *Pulse Information Of Broadcasting Process*

When the system scheduling service process is running, the pulse information is broadcast every 1 second. If the system monitoring service process does not receive pulse information, the system monitoring service process restarts the system scheduling service process according to the election algorithm.

C. System Monitoring Service Process

Handle system scheduling process commands. Command processor scan command manager. If there is an unprocessed command, read the command and parse it, then execute the command and send the execution results to the system scheduling service process through the TCP communication module; if the command manager is empty, then wait for 1 millisecond to continue scanning [14] [15].

The process monitoring subsystem is mainly composed of TCP communication module, command manager and command processor. TCP communication module is responsible for communication with process scheduling subsystem; command processor is used to save system maintenance client operation commands; command processor is used to process system maintenance client operating commands.

1) Receiving System Scheduling Service Process Commands

The operation command of the receiving process system scheduling reset process is completed by an independent thread. System monitoring service process and system scheduling service process communicate through TCP. The TCP communication module of the system monitoring service process is responsible for receiving the operation commands of the system dispatching service process. System scheduling service process operation commands are received and written to the command manager.

2) Handling System Scheduling Service Process Operation Command

The command processor is completed by independent threads. Command processor scan command manager. If there is an unprocessed command, read the command and parse it, then execute the command and send the execution results to the system scheduling service process through the TCP communication module; if the command manager is empty, then wait for 1 millisecond to continue scanning.

3) Start the System Scheduling Service Process

The system monitoring service process starts with the operating system, and then starts the system scheduling service process according to the ' election ' algorithm.

4) Pulse Information of Broadcasting Process

According to the design requirements, the target process broadcasts pulse information outward UDP every second. The definition of pulse information includes: host unique identification, host IP, TCP port, process identification, running identification and other information.

IV Conclusion

Multi-inference real-time monitoring software has the advantages of real-time monitoring, automatic scheduling and load balancing, which can make the monitored inference process run stably and reliably for a long time without anybody on duty. The author believes that the system can not only monitor the running state of multiple inference engines in real time, but also be widely used in many occasions that require process monitoring. At present, there is no domestic application in this area, it can be said that the product to fill this gap.

Acknowledgment

Fund Project: Supported by the Special Project of Shaanxi Provincial Department of Education (No.: 17JK0388).

References

[1] Gaoli, Bu Huaiyuan, Hushu. A medical diagnosis reasoning machine design and implementation [J]. Computer application and software, 2002 (6): 44-46.

[2] Gu Shenming, Liu QuanLiang.An expert system is based on Web design and implementation [J]. Computer engineering. 2001, 27 (11): 100-101.

[3] ZhaoZidu.Reasoning mechanism and reasoning method [J]. Automation expo, 1997 (6): 19-20.

[4] Wu Quanyun, etc. Artificial intelligence and expert system [M]. Anhui: national defense science and technology university press, 1995, 168-172.

[5] ChenZhaoQian, et al. DBEST: apractical fault diagnosis expert System development tools [J]. Microcomputer, 1995, 15 (6): 16-20.

[6] Guo Huawei, Shi WenKang, DengYong, et al. Conflict of evidence:discard, discover or to solve [J]. Systems Engineering and electronics, 2007, 29 (6): 890-898.

[7] PaksoyA, Gkt ürkM.Information fusion with dempster-shafer evidence theory for software defect prediction [J]. Procedia Computer Science, 2011 (3): 600-605.

[8] Hartley RVL.Transmission of information [J]. Bell Systems Technical Journal, 1928, 7 (3): 535-563.

[9] JousselmeAL, GrenierD, Bosse. A new distance between Two bodies of evidence [J]. InformationFusion, 2001, 2 (2): 91 – 101.

[10] 5th National Conference on Artificial Intelligence, Philadel- phia, USA, Aug. 11 – 15: 869-901.

[11] Honsel Daniel, Herbold Verena, Waack Stephan et al. Investigation and prediction of open source software evolution using automated parameter mining for agent-based simulation [J] Automated Software Engineering, 2021, 28(1).

[12] Yinling Liu, Tao Wang, Haiqing Zhang et al. An improved approach on the model checking for an agent-based simulation system [J] Software and Systems Modeling, 2020.

[13] Chouaki Tarek, Puchinger Jakob Agent based simulation for the design of a mobility service in the Paris-Saclay area [J] Transportation Research Procedia, 2021, 52.

[14] Feather Christopher Footing the Reconstruction Bill: An Appraisal of the Financial Architecture for Disaster Rebuilding in the United States of America [J] International Journal of Disaster Risk Reduction, 2021.

[15] Bi Tingting, Ding Wei, Liang Peng et al. Architecture information communication in two OSS projects: The why, who, when, and what [J] The Journal of Systems & Software, 2021, 181.

Design and Implementation of Intelligent Agricultural Greenhouse System

Kaifa Kang
Dept of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: ttla02@126.com

Lei Tian
School of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: tianlei@xupt.edu.cn

Qingmin Zhang
School of Communication Engineering
Xidian University
Xi'an, China
E-mail: qmzhang@stu.xidian.edu.cn

Xu Yanrui
School of Humanities and foreign languages
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: 364945696@qq.com

ShuKang Wei
School of Electronic Engineering
XUPT
Xi'an, China
E-mail: 2397318147@qq.com

Abstract—With the rapid development of science and technology, it has brought about the improvement of human social productivity. It is very necessary to apply embedded technology to agriculture. Smart agriculture will gradually replace traditional agriculture, liberate productivity, and promote agricultural development more efficiently and intelligently. In recent years, intelligent agricultural greenhouse system has increasingly appeared in people's vision and played a great role in agriculture. At present, there are many intelligent agricultural greenhouse systems, but they have single function, relatively low integration and limited application environment. They are only suitable for large-scale cultivated land. Based on this, this paper designs and implements a more efficient and widely used intelligent agricultural greenhouse system, which plays a role in the development of agricultural greenhouse and better applies embedded technology to agricultural development. The system mainly completes three parts of functions: the first part uses the temperature and humidity module and light intensity module to display the collected temperature and humidity information and light intensity information on the screen through the STM32 single chip microcomputer FSMC interface, and upload these information to the terminal (upper computer) through

DOI: 10.21307/ijanmc-2021-036

the ESP8266 module for real-time viewing by the greenhouse owner. The second part is to automatically supplement the light source according to the change of light intensity through STM32 single chip microcomputer to promote the growth of crops in the greenhouse. The third part is to realize the function of monitoring and recording.

Keywords-STM32; Intelligent Agricultural Greenhouse; Light Intensity Adjustment; Sensors; Monitor

I. INTRODUCTION

In recent years, with the rapid development of China's industry, intelligent agricultural greenhouses have introduced advanced modern industrial technology. The farming methods in rural areas have been greatly improved from the original pure human labor to today's semi mechanized farming. Farmers can work more labor-saving and obtain more benefits at the same time. China is a large agricultural country. It can be learned from the seventh national census that China's current population is about 1.41 billion. Therefore, China has great pressure on agriculture. Through continuous research and development, the hardware quality level and supporting capacity of intelligent agricultural greenhouses are improved [1].

According to the status of the development of China's intelligent agricultural greenhouse and China's agricultural personalized form of expression, have a higher degree of integration, a wider range of application of intelligent agricultural greenhouse system is more conducive to promote the development of China's intelligent agricultural greenhouse [2-3].

Intelligent agriculture has created opportunities for the development of agriculture in China. The intelligent agricultural technology will rely on the interconnection of all things under 5G to diagnose the crop growth environment and growth conditions, such as detecting the environmental temperature, humidity, light intensity and other information, and then put forward strategies to make the intelligent agricultural system adjust a more appropriate temperature and let unmanned aerial vehicles spray fertilizers and pesticides.

II. OVERALL SCHEME DESIGN

The STM32F407 system uses DHT11 module, 2 million-pixel OV2640 module, photosensitive sensor, 4.3-inch TFTLCD screen, SD card, ATK-ESP8266 serial port to WiFi module, infrared detection module, light source adjustment circuit and buzzer in hardware [4].

Under normal conditions, DHT11 collects temperature and humidity, the photosensitive sensor collects light intensity, updates the display on the LCD screen in real time, and sends the data to the server through serial port 3 to WiFi [5]. When it is detected that the light intensity is less than a certain threshold, the light source adjustment circuit is called to automatically adjust the light source to improve the light intensity. When other people or animals break into the greenhouse, the infrared module monitors, transmits the detected signal to the STM32 development board, which responds, calls the active buzzer for alarm, OV2640 captures this image for storage, displays the latest captured image on the LCD screen, and finally sends the alarm signal to the server through serial port 3 to WiFi [6].

III. SYSTEM HARDWARE DESIGN

A. Information Collection Module

1) DHT11sensor

DHT11 sensor is a temperature and humidity composite sensor with calibrated digital signal output [7]. Its features include low cost, good stability, can also be used in harsh environment, fast response to these aspects. DHT11 adopts the serial mechanism. The circuit design of DHT11 module is shown in Figure 1.

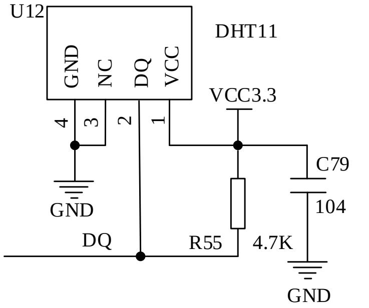


Figure 1. Circuit design of DHT11 module

2) Optical intensity module

This part uses the photosensitive sensor to form the light intensity module. Photosensitive sensors convert the optical signals into electrical signals by using the characteristics of photosensitive elements, and their response wavelengths are near visible wavelengths.

Generally, when detecting light intensity, photosensitive sensor is also often used as a device. If other signals are converted into optical signals and then through the photosensitive sensor, the parameter values of other signals can also be obtained. The photosensitive sensor used in this design is a photosensitive diode, whose structure is similar to the semiconductor structure, with a PN junction inside, so the photosensitive sensor also has one-way conductivity. Schematic diagram of optical intensity module is shown in Figure 2. LS1 is a photodiode and R58 provides it with a reverse voltage. When the light intensity changes, the voltage at both ends of LS1 will also change. The stronger the light intensity is, the lower the corresponding voltage will be; otherwise, the higher the light intensity is, the higher the corresponding voltage will be. The LIGHT SENSOR connects to the PF7 pin of the development board and uses the analog input function of the pin. At the same time, the onboard ADC3 channel is used to convert the collected analog signal into digital signal.

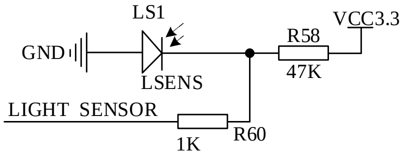


Figure 2. Schematic diagram of optical intensity module

3) Infrared detection module

Infrared detection module HC-SR501 is an automatic control module using infrared technology [8]. This module works in the dc voltage range of 4.5-20V and can output TTL level. It has a working range of 7 m. It has two triggering modes: Non-triggering mode and repeatable trigger mode.

This module has inductive blocking time, that is, when the inductive output high level becomes low level, the developer can manually set a blocking time, during the blocking time, the infrared sensor does not receive any inductive signal. Therefore, it plays an important role in interval detection and effective suppression of various interferences in the process of load switching.

4) Camera module

The OV2640 offers a single UXGA camera. Generally, OV2640 module is controlled by SCCB bus. The OV2640 can achieve a maximum frame rate of 15 frames per second if in output UXGA image mode. If through SCCB interface programming can achieve any image processing process [9]. The OV2640 module consists of the following modules: photosensitive array, analog signal processing, 10-bit A/D conversion, 8-bit microprocessor [10].

B. Information Transmission and Display

1) ESP8266 module

Because ESP8266 module supports TTL serial port, if the TTL level of MCU is 3.3V or 5V, this module can be used. After firmware burning, the module can support three connection modes to realize data transmission, which are: serial port to WIFI STA mode, serial port to AP mode, serial port to STA+AP mode [11].

2) TFTLCD screen

This part is mainly used for the display of the intelligent agricultural greenhouse system. It can refresh the information of temperature, humidity and light intensity in real time under normal circumstances, display the normal operation of the camera module during the capture, capture the picture and keep the latest capture on the screen.

TFTLCD screen is different from the simple matrix of passive TNLCD and STNLCD. Each pixel of TFTLCD corresponds to a thin film transistor. Therefore, the crosstalk of TFTLCD screen can be eliminated, and the static characteristics of the LCD screen can be independent of the number of scan lines, which greatly improves the image quality of the display.

C. Information Storage and Processing

1) SD module

This module is mainly used to store images taken. Considering to support SPI/SDIO driver, and large storage capacity, THE selection of SD card, and its MCU system used to do external memory is very convenient. Use the development board PC8, PC9, PC10, PC11, PC12, PD2 pins [12-13].

The SDIO interface has the advantage of being compatible with various memory cards. The SDIO controller of STM32F4 consists of SDIO adapter module and APB2 bus interface connecting CPU. The block diagram of its functional parts is shown in Figure 3. In general, SDIO_D0 is used for data transfer, and after initialization the host can be used to change the width of the data bus. If the memory card of multimedia type is connected to the bus, SDIO_D0, SDIO_D[3:0] or SDIO_DO, SDIO_D[7:0] can be used for data transmission. If the SD card is connected to the bus, the host is configured with SDIO_D0/SDIO_D[3:0] for data transmission. If you initialize, SDIO_CMD needs to be set to open mode; For command transfer, SDIO_CMD needs to be set to push-pull mode.

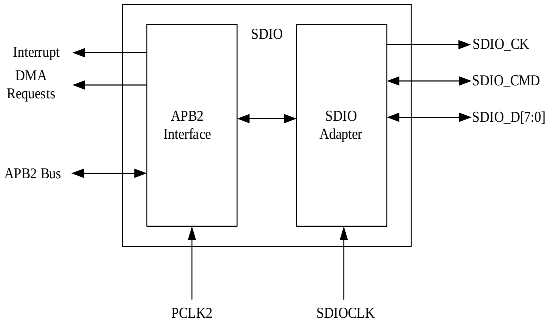


Figure 3. Circuit diagram of light source module

2) Light source module:

The light source module automatically adds light intensity according to the change of light intensity, and is connected with the PA2 pin of the development board for multiplexing function. Multiplexing is the output comparison function of timer 9 to control the intensity of light source. [14] This part of external circuit is composed of light-emitting diode in series with a pull-up resistor, and its circuit diagram is shown in Fig. 4.

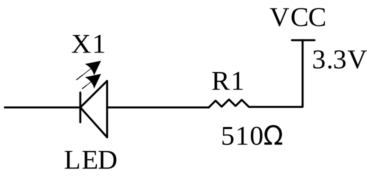


Figure 4. Light source module circuit diagram

3) Alarm module

The alarm module mainly uses STM32 to control the buzzer to send sound alarm and WiFi module to send alarm signal to the server. For details about how to send alarm signals, see 3.2.1ESP8266. The buzzer alarm is also relatively simple. It is connected to the PF8 pin of the development board for general output function, and its schematic diagram is shown in Figure 5.

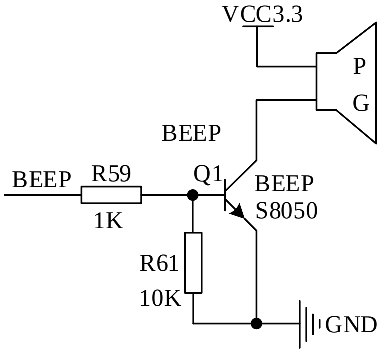


Figure 5. Schematic diagram of buzzer alarm module

IV. System software design

The system software design mainly introduces the code analysis of each function, including initialization setting of hardware, code analysis of abstraction layer, analysis of important functions and code explanation of specific functions. The following is mainly a flow chart + description.

A. Collection, Display and Transmission

1) Collection of the data

Write the temperature and humidity collection code dht11.c according to the working time sequence diagram of DHT11 temperature and humidity sensor. DHT11_Rst() is used to reset the DHT11, DHT11_Check() is used to wait for the DHT11 response, and DHT11_Read_Data() is used to read the temperature and humidity data. In addition, it is also important to initialize the PG9 pin to pull up, normal output mode.

Lsens.c and adc3.c files are written according to the working principle of photosensitive sensor to complete the acquisition of light intensity. [15] Since ADC channel is used here to obtain the analog value, the configuration function of ADC initialization is first written, and the function of AD converter Adc3_Init() is turned on. Then write the Get_Adc3() function to obtain the value of the seventh channel of ADC3, then initialize the photosensitive sensor to set PA7 as the analog input function, and write Lsens_Get_Val() to complete the acquisition of light intensity value.

After writing functional functions, we can call DHT11_Read_Data() and Lsens_Get_Val() to complete the collection of temperature, humidity and light intensity.

2) Display of the data

Temperature and humidity light intensity display will mainly use TFTLCD screen, here need to write the screen driver and Chinese characters display function.lcd.c is used to drive the screen and text.c is used to display Chinese characters.

First, write lcd.c according to LCD usage flow chart (Fig. 6). Initialize the required pins to drive the FSMC interface, initialize the LCD screen by using the initialization series in the screen's data manual, and finally set the coordinates and write GRAM to display characters or numbers.

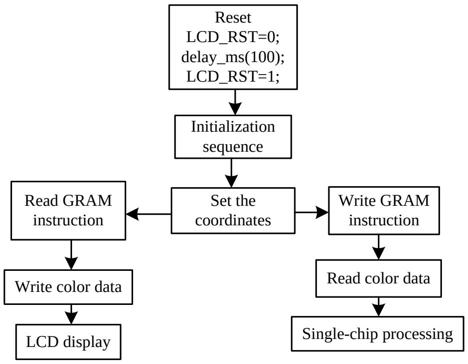


Figure 6. LCD usage flow chart

Text.c is written to make it easier to display text content, Get_HzMat() is used to find a font from a font library, Show_Font() is used to display characters of a specified size, and Show_Str() is used to display a string.

At the same time, in order to update the temperature and humidity and light intensity in real time without interfering the operation of the main program, timer 4 is started in this design, and TIM4_Int_Init() is written to initialize timer 4 and enable the NVIC interruption of timer 4.

Write timer 4 interrupt service program TIM4_IRQHandler() to obtain real-time temperature and humidity, light intensity information in real time and display it on LCD screen.

Show_Str() and LCD_ShowNum() are used to display the temperature, humidity, and light intensity. Their code in the program is as follows:

```
Show_Str(30,90,400,24,"LSENS_VAL:",24,0);

LCD_ShowNum(30+10*12,90,adcx,3,24); //Displays light intensity

Show_Str(30,130,300,24,"Temp:    C",24,0);

LCD_ShowNum(30+60,130,temperature,2,24);  //Display the temperature

Show_Str(30,170,300,24,"Humi:    %",24,0);


LCD_ShowNum(30+60,170,humidity,2,24); //Display the humidity
```

3) Transmission of the data

Transmission is the key and difficult point in this design. In this part, esp8266. c is mainly written to establish the connection, usart.c is used to send data through the serial port. In short, this part is to control WiFi module to send and receive data through TCP/IP protocol in wireless local area network through AT instruction set.

In common AT instruction sets, "AT+WAMODE=Y" indicates setting the working mode of WiFi module. At present, there are three working modes of WiFi module, which are as follows: Y equals 1 indicates that the module is in STA mode, and the function of the module is to connect to the wireless network as a wireless WiFi STA. Y = 2 means that the WiFi module is set in AP mode. At this time, the module can act as a WiFi hot spot and allow other WiFi to connect to the module. Y = 3 means that the WiFi module is set to STA+AP mode. At this time, the WIFI module has two functions at the same time. It can be used as a hot spot for the connection of other WiFi devices, and can also be added to other WiFi hot spots. One pattern corresponds to three sub-patterns. This design uses STA mode in which the module connects to other WiFi hot spots. Next, the three modes under STA are introduced in detail, and the sub-modes of the other two modes are similar. The WiFi module can be set as TCP server, TCP client and UDP in STA mode, and the corresponding upper computer is set as TCP client, TCP server and UDP respectively. In this design, WiFi module in STA mode is used as TCP client and configured as TCP server in the upper computer software. Table 4.1 shows the configuration in TCP client mode.

After mastering the AT instruction, it is necessary to write the WiFi module function, write esp8266_start_trans () to connect the hot spot and establish the connection with the server software of the upper computer, send data to the upper computer using esp8266_send_data (), Exit transparent transmission and close the connection with the server through the esp8266_quit_trans() function after the data is sent.

As the serial port to WiFi module is used, the process of data transmission is that MCU sends data to serial port 3, and then serial port 3 receives data and sends it to the upper computer server through WiFi module.

The function of using serial port 3 also needs to be configured. Write usart3_init() with the corresponding pin initialization and NVIC interrupt Settings.

To ensure that the transmission is not disturbed, the NVIC interrupt here should be set to have a high preemption priority. To ensure the correctness of received data, the serial port needs to invoke timer 7 to initialize the serial port, and use the timer 7 interrupt service function to clear data before receiving. Set the interrupt service function USART3_IRQHandler() to receive data after serial port 3 is initialized. Write the u3_printf() function to send data, which is used in writing the esp8266.c data send function, or to send data directly after the ESP8266 module has established a connection to the server.

TABLE I. STA AND TCP CLIENT CONFIGURATION

Instructions	Functions
AT+CWMODE=1	Set the WiFi module to STA mode
AT+RST	Restart the WiFi module to take effect
AT+CWJAP="xx","xxxxxxx"	Add a WiFi hot spot: xx, the password is xxxxxxxx
AT+CIPMUX=0	Enabling single connection
AT+CIPSTART="TCP","192.168 .1.XXX",8000	Establish a TCP connection to 192.168.1.xxx,8000
AT+CIPMODE=1	Enable transparent transmission
AT+CIPSEND	Begin to transport

B. Monitoring and Storage

1) Infrared detection and alarm

This part of the function is mainly realized by writing dt.c. The PA3 pin is first configured to initialize it, and the PA3 pin is set to normal input and drop-down mode. Through PA3 pin to infrared detection module output level detection, detection of rising edge trigger interrupt. External interrupts are implemented on interrupt line 3 and need to be configured for interrupt initialization. Configuration interrupt initialization includes initializing PA3, enabling interrupt clock configuration interrupt, connecting PA3 pins to interrupt line 3, and configuring NVIC interrupts for interrupt line 3 and interrupt line 3.

2) Capture and storage

This part of the software is designed to capture and store images. malloc.c is the memory management driver code. The FATFS file system code is open source and only needs to be ported for use on the development board. sdio_sdcard.c is the code for storing data to an SD card. sccb.c, ov2640.c, and photo.c are used to capture JPG images. dcmi.c is used to transfer captured image data to SD card. The process flow chart of capturing process is shown in Figure 7.

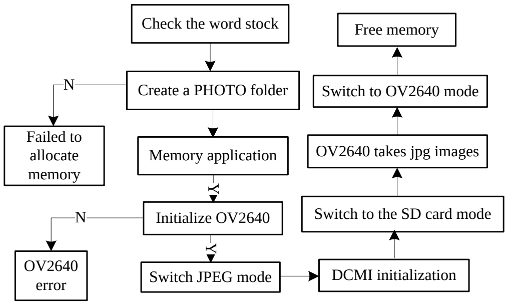


Figure 7. Flowchart of program realization of capture

To accomplish this function, the driver code of SCCB interface should be written first. The second step is to write the OV2640.c code to turn on the OV2640 module. The OV2640 uses OV_SCL and OV_SDA to configure registers, as well as signals such as OV_PWDN and OV_RESET. Configure OV2640 initialization as shown in Figure 8. Another important function here is to set the image output window function

OV2640_Window_Set(), set the image output size function OV2640_OutSize_Set(), set the window function OV2640_ImageWin_Set(), set the image resolution size function OV2640_ImageSize_Set ().

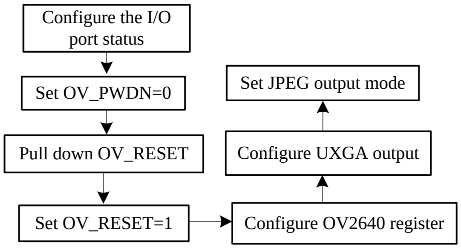


Figure 8. OV2640 initial configuration flowchart

The third step is to compile dcmi.c which mainly completes four functions. The first is to enable the clock, configure the mode of the required pins and set the reuse function. Then complete the configuration of the DCMI. In this step, important parameters such as HSPOL/PCKPOL/VSPOL data width in the DCMI_CR register need to be configured. When frame interrupt is enabled, DCMI interrupt service function is written for data processing.

3) Automatic Adjustment of Light Intensity

In the program prepared pwm.c to control the intensity of the light source to achieve this function. Initialize timer 9PWM, define the structure variable required by timer 9, enable TIM9 clock and port A clock, reuse PA2 pin to timer 9, initialize PA2 pin, and then initialize timer 9, because the PWM mode of timer 9 channel 1 is used here, Therefore, it is very important to initialize the channel. Finally, enable timer 9 to complete the setting of timer 9PWM wave.

Then we set the comparison value of PWM wave according to the obtained light intensity value, so as to modify the duty cycle to achieve the purpose of adjusting the light source. The larger the intensity value, the smaller the comparison value and the weaker the light source. This process is shown in Table 2.

TABLE II. AUTOMATIC ADJUSTMENT OF LIGHT SOURCE

Light intensity values (adcx)	**Comparing values (ledpwm val)**
adcx＞50	0
30＜adcx≤50	100
20＜adcx≤30	300
adcx≤20	500

V. FUNCTION IMPLEMENTATION AND TESTING

The physical picture of the system is shown in Figure 9. And the system startup interface is shown in Figure 10. After entering the startup interface, you can get the function introduction of the system, the name of the system, the model of the development board used, and the department of the system.

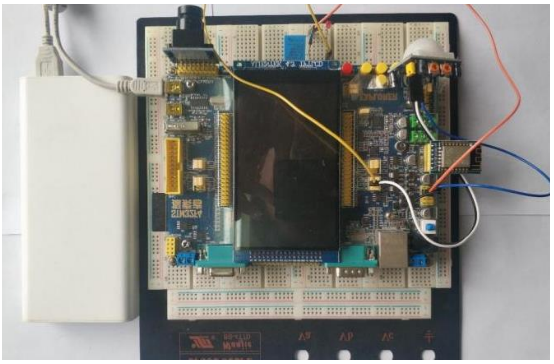
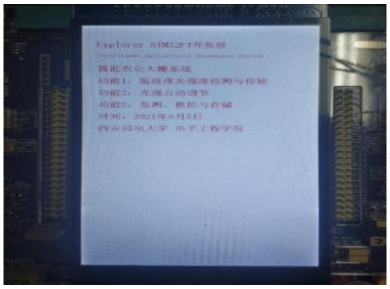


Figure 9. System physical drawing

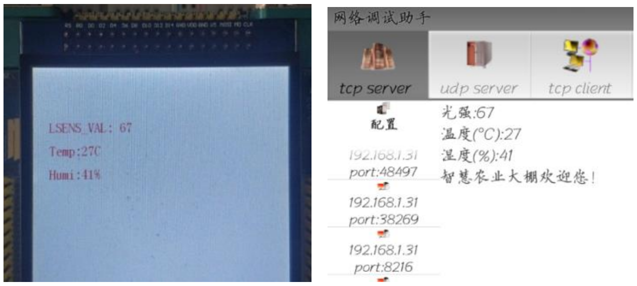


Explorer STM32F4开发板
Intelligent Agricultural Greenhouse System
智能农业大棚系统
功能1：温湿度光强度检测与传输
功能2：光强自动调节
功能3：监测、抓拍与存储
时间：2021年6月5日
西安邮电大学 电子工程学院

Figure 10. System startup interface

A. Measurement of the Data

The test here is mainly on the screen and the upper computer temperature and humidity size, light intensity size test. The test environment was 8 a.m., 13 p.m., and 20 p.m. Figure 11 (a) and Figure 11 (b) respectively show the display of temperature, humidity and light intensity on LCD screen and upper computer at 8:00 in the morning. Table 3 shows the measured temperature, humidity and light intensity.



192.168.1.31
port:48497
192.168.1.31
port:38269
192.168.1.31
port:8216

光强:67
温度(℃):27
湿度(%):41
智慧农业大棚欢迎您!

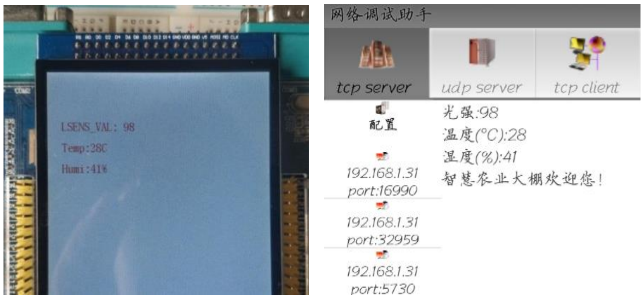
(a) (b)

Figure 11. (a) Screen display at 8:00, (b) Upper computer display at 8:00

TABLE III. TEST RESULTS AT 8:00

Time	Light intensity	Temperature	Humidity
8:00	67	27℃	41%

Figure 12 (a) and (b) respectively show the display of temperature, humidity and light intensity by LCD screen and upper computer at 13:00 at noon. Table 4 shows the specific temperature, humidity and light intensity measured at this time.



192.168.1.31
port:16990
192.168.1.31
port:32959
192.168.1.31
port:5730

光强:98
温度(℃):28
湿度(%):41
智慧农业大棚欢迎您!

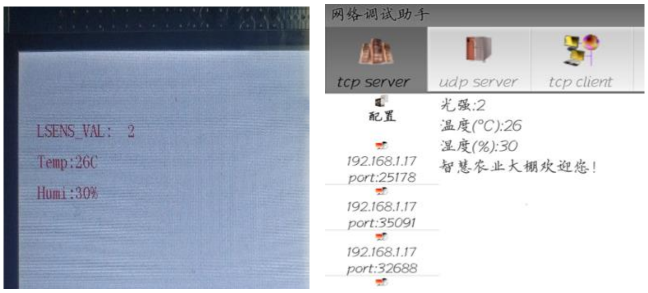
(a) (b)

Figure 12. (a) Screen display at 13:00; (b) Display of upper computer at 13:00

TABLE IV. TEST RESULTS AT 13:00

Time	Light intensity	Temperature	Humidity
13:00	98	28℃	41%

Figure 13 (a) and (b) respectively show the display of temperature, humidity and light intensity by LCD screen and upper computer at 20:00. Table 5. Shows the specific temperature, humidity and light intensity measured at this time.



192.168.1.17
port:25178
192.168.1.17
port:35091
192.168.1.17
port:32688

光强:2
温度(℃):26
湿度(%):30
智慧农业大棚欢迎您!

(a) (b)

Figure 13. (a) Screen display at 20:00; (b) Upper computer display at 20:00

TABLE V. TTEST RESULTS AT 20:00

Time	Light intensity	Temperature	Humidity
20:00	2	26℃	30%

The tests in three time periods verified the correctness of temperature, humidity and light intensity test results, as well as the correctness of data display on LCD screen and transmission through ESP8266WiFi module.

B. Automatic Adjustment of Light Source Measurement

In this part, we observe the light source module by changing the light intensity. Under normal circumstances, the light source is usually warm color light, but here, in order to make the test results more obvious, red LED is selected for testing. In the first set of data, if the light intensity is greater than 50, it can be seen that the light source does not emit light, as shown in Figure 14.

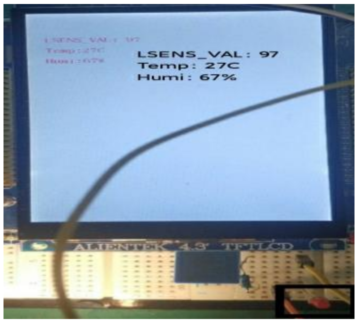


Figure 14. The first group of auto-adjusting light source

The second set of data makes the light intensity within the range of 20-30, and it can be seen that the light source is shining, but not particularly bright, as shown in Figure 15.

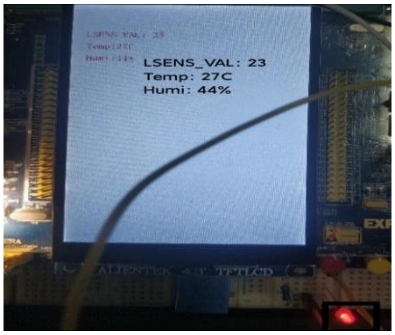


Figure 15. The second group of test pictures of automatic adjusting light source

The third group of data makes the light intensity less than 20. It can be seen that the light source is the brightest at this time, as shown in Figure 16. The light source module is located at the lower right corner of Figure 14-16.

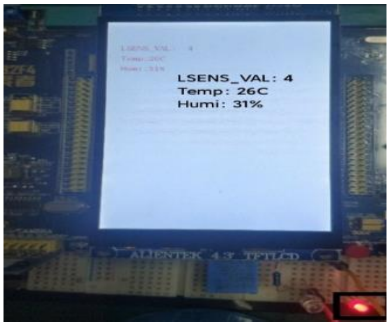


Figure 16. The third group of test pictures of automatic adjusting light source

At this point, the automatic adjustment of the light source to complete the test, the test results are normal.

C. Infrared Monitoring

This part of the function needs to detect animals or people can be captured. In this part, two groups of tests were conducted, each of which tested the following five indicators:

1) *Captured image information;*

2) *Whether the system can return to normal state after capturing (that is, temperature, humidity and light intensity are displayed and the next capturing can be carried out);*

3) *Keep the latest snapshot images on the screen;*

4) *Verify whether the captured image is stored on the SD card;*

5) *Whether the buzzer sends sound alarm and whether the upper computer receives the alarm signal "Warning!".*

According to these 5 indicators, the first group of images obtained after testing are shown in Figure 17-20.

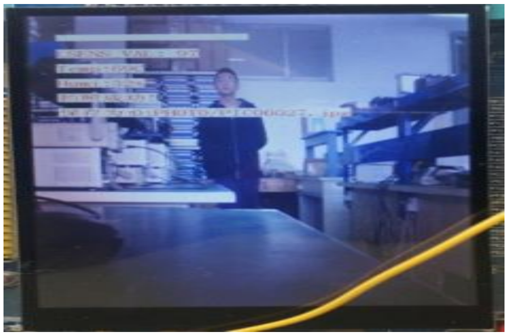


Figure 17. Capture group 1 (1)

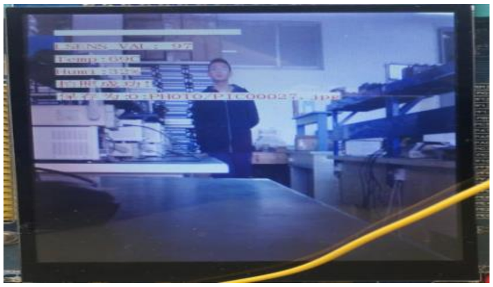


Figure 18. Capture group 1 (2)

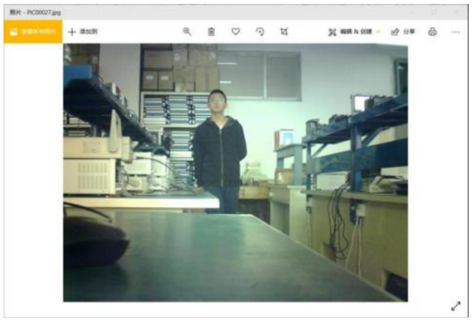
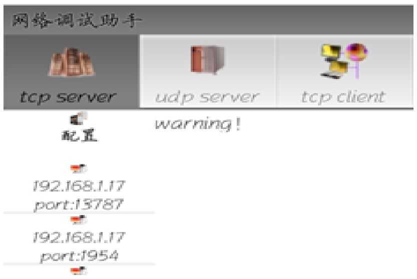


Figure 19. Capture group 1 (3)



warning!

192.168.1.17
port:13787

192.168.1.17
port:1954

Figure 20. Capture group 1 (4)

Figure 17 is the original image captured. Figure 18 shows the image of screen retention and return to normal state after capturing. From here, we can also get the location information of the saved image after capturing: PHOTO/PIC00027.jpg; Figure 19 shows the image information stored on the SD card. The realization of capturing and storage functions can be verified through Figure 19 and Figure 20 is the alarm signal sent to the upward machine after monitoring.

Figure 21-24 shows capturing the second group of images.

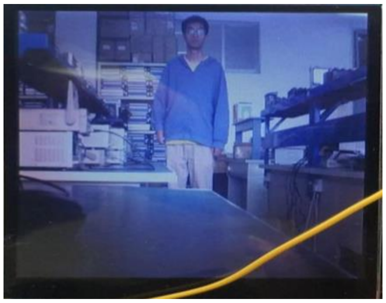


Figure 21. Capture group 2 (1)

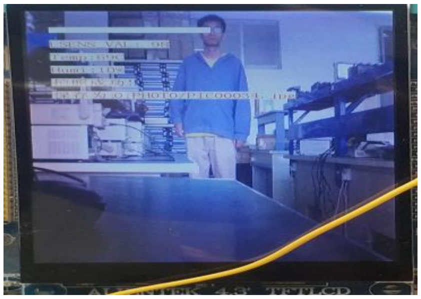


Figure 22. Capture group 2 (2)

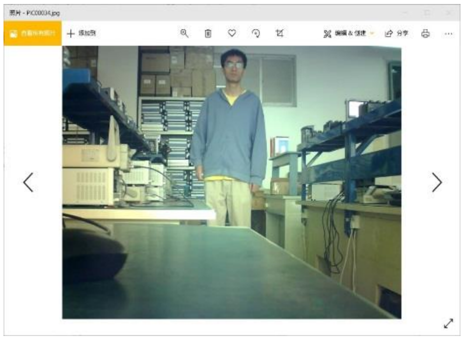
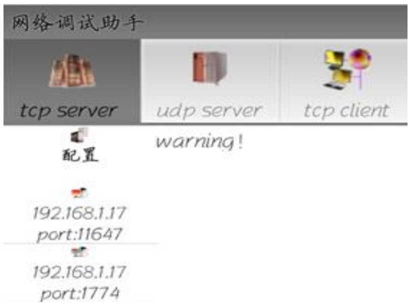


Figure 23. Capture group 2 (3)



warning!

192.168.1.17
port:11647

192.168.1.17
port:1774

Figure 24. Capture group 2 (4)

The second set of tests was conducted mainly for further verification. Figure 21 shows the original image captured for the second time. Figure 22 shows the image of screen retention and return to normal state after the second capture. From here, we can also get the location information of the saved image after capturing: PHOTO/PIC00034.jpg; Figure 23 shows the image information stored on the SD card. The realization of capturing and storing functions can be verified again through Figure 23 and Figure 24 is the alarm signal sent to the computer again after infrared detection.

By comparing the two captured images with the test indexes, the infrared monitoring alarm and storage functions have been verified.

From here, we can also get the location information of the saved image after capturing: PHOTO/PIC00027.jpg; By comparing the two captured images with the test indexes, the infrared monitoring alarm and storage functions have been verified.

ACKNOWLEDGMENT

This work was partly supported by the Natural Science Basic Research Program of Shaanxi (Program No.2021JM-460) and the Scientific Research Program Funded by Shaanxi Provincial Education Department (No.21JC033); the higher education scientific research project of Shaanxi higher education society (No. XGH21161) and the undergraduate innovation and entrepreneurship plan (No.202111664039, S202111664070 and 202111664019).

REFERENCES

[1] J. Chao and E. Steinbach, “Preserving SIFT features in JPEG-encoded images,” 2011 18th IEEE International Conference on Image Processing, 2011, pp. 301-304, doi: 10.1109/ICIP.2011.6116299.

[2] Viswanathan, S. , et al. “A model for the assessment of energy-efficient smart street lighting-a case study.” Energy Efficiency, vol. 14, Jun. 2021, pp.1-20, doi:10.1007/s12053-021-09957-w.

[3] Z.Q. Wang, W. Huang, L Tong, et al. “Design of Timing Charging and Discharging System for Pneumatic or Hydraulic Pressure Device Based on STM32.” In Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering (EITCE 2020). Association for Computing Machinery, NY, USA, 1108–1112. doi:10.1145/3443467.3443913

[4] Liu, C., Q. Wang, and F. Zhang. “Design and development of city street-lighting energy-saving system.” IEEE, Aug. 2010, doi:10.1109/PACCS.2010.5627060.

[5] S. Matlak and R. Bogdan, “Reducing energy consumption in home automation based on STM32F407 microcontroller,” 24th Telecommunications Forum (TELFOR), 2016, pp. 1-4, doi: 10.1109/TELFOR.2016.7818776.

[6] Yasuhisa Omura, “Concept of an Ideal pn Junction,” Bipolar-type Insulated-gate Transistors , IEEE, 2013, pp.1-5, Doi: 10.1002/9781118487914.ch01.

[7] Ping, H., and C. Tang. “Indoor detection based on MLX90621 infrared sensor.” Electronic Measurement Technology, vol. 39, Aug. 2016, pp.118-121, doi: 10.19651/j.cnki.emt.2016.08.025.

[8] Yan, A. I., and G. University. “Intelligent Lighting Control System PWM Dimming Smoothing Optimization.” Times Agricultural Machinery, vol. 44, May. 2017, pp.114-115.

[9] JJ Sáenz-Peafiel, J. L. Poza-Lujan, and JL Posadas-Yagüe. “Smart Cities: A Taxonomy for the Efficient Management of Lighting in Unpredicted Environments.” DCAI, Jun. 2019, pp.63-70.

[10] Sun, F., and J. Yu. “Indoor intelligent lighting control method based on distributed multi-agent framework.” Optik-International Journal for Light and Electron Optics, vol 213, Jul. 2020, pp.1-10, doi:10.1016/j.ijleo.2020.164816.

[11] Piao S, Ciais P, Huang Y, et al. “The impacts of climate change on water resources and agriculture in China.” Nature, vol. 467, Jul. 2010, pp.43-51.

[12] Tolomio M, Casa R. “Dynamic crop models and remote sensing irrigation decision support systems: A review of water stress concepts for improved estimation of water requirements.” Remote Sensing, vol. 12, 2020, pp. 3945. Doi:10.3390/rs12233945.

[13] Jose M G, Pereira LS. “Decision support system for surface irrigation design.” Journal of Irrigation & Drainage Engineering, vol. 135, 2009, pp. 343-356. doi: 61/(ASCE)IR.1943-4774.0000004.

[14] Dadari S and Ahmadi S H. “Calibration and evaluation of the FAO56-Penman-Monteith, radiation, and Priestly-Taylor reference evapotranspiration models using the spatially measured solar radiation across a large arid and semi-arid area in southern Iran.” Theoretical & Applied Climatology, vol. 136, 2019, pp. 441-455. doi: 10.11126/0204112877186.

[15] Sun Y P, Lan Y P. “Research on self-learning fuzzy control of controllable excitation magnetic suspension linear synchronous motor.” Journal of Electrical Engineering and Technology, vol. 15, 2020, pp. 843-854. doi: 10.1007/s42835-020-00347-3.

Feature Sorting Algorithm Based on XGBoost and MIC Combination Model

Gao Xiang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 971104101@qq.com

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

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

Hu Yuzhe
Jinan University-University of Birmingham Joint Institute
Jinan University
Guangzhou, 511400, Guangdong, China
mail:18137910896@163.com

***Abstract*—Feature ranking can not only help the data analysis system improve efficiency, but also reduce the interference of redundant features and irrelevant features to the results. At present, feature ranking of massive data is an important and difficult problem. In order to solve the above problems, this paper proposes a feature importance ranking algorithm based on XGBoost and MIC model by analyzing the existing algorithm models. Firstly, XGBoost model and MIC model are established respectively; Then, the results of the above two models are weighted and combined by the error reciprocal method. XGBoost model has the advantages of high efficiency, flexibility and portability, while MIC model has universality and easy parameter adjustment. The resulting XGBoost MIC combination model has both advantages; Finally, the first mock exam is used as a sample set of data for anticancer drug candidates. After preprocessing the data set, the XGBoost-MIC combination model is used to analyze the case. At the same time, the calculation results of a single model are calculated, and the model is optimized by adjusting the parameters of the model. The results show that the error of the first mock exam is obviously lower than that of the single calculation model, and the accuracy of the XGBoost-MIC is 0.75, which is 0.02 higher than that of the single model.**

Keywords-Feature Sorting; MIC Arithmetic; Xgboost Arithmetic; Combination Model

I. INTRODUCTION

With the rapid development of the Internet, the amount of data generated by human activities is increasing exponentially. Big data is generally considered as PB-level data, including structured, semi-structured and unstructured data, whose scale and complexity greatly exceed the storage and computing capacity of existing hardware. However, due to the variety of irregular or incomplete data, such data brings great difficulties to data processing. Moreover, the traditional GBDT has defects in design, can not be processed in parallel, has high computational complexity, and is not suitable for high-dimensional sparse features.

In order to tackle a series of difficulties and challenges brought by massive data to server storage and operation, researchers have been trying to find out an efficient data analysis system that can extract valuable information from massive data. In recent years, many Internet companies have launched various big data processing systems, such as Google's MapReduce system, Microsoft's Cosmos system developed for parallel Blockchain, Haystack system proposed by Facebook to solve

DOI: 10.21307/ijanmc-2021-037

the massive small files, etc. Under the background of massive data, it is more important to grasp the key information of data quickly. The big data application, which is named “ScholarSpace”, developed by the Network and Mobile Data Management Laboratory (WAMDM) of Renmin University of China fully reflects the process of big data processing, and combines modern big data technology with traditional economics, law, literature and other disciplines to obtain key information. However, because the big data is not structured and its relationship is complex, it is not easy to establish association between the data object’s structure and attributes. Moreover, because the error value of the single model for calculating the importance is high, the algorithm for extracting key information from big data always has certain limitations. In addition, other researchers also began to pay attention to the importance of feature selection system in massive data. Based on LDA model, Yang Guijun quantified the key information as subject feature vector as explanatory variable, and integrated multiple models with sample disturbance and attribute disturbance by using XGBoost algorithm to build prediction model; According to the film review data, Zhang Hongli extracted features from user reviews as auxiliary prediction indicators, and combined with other factor indicators as independent variables to build a prediction model; Based on the thought process of constructing tree in XGBoost algorithm, Li Zhanshan proposes a new wrapped feature selection algorithm xgbsfs, which avoids the limitation of single importance measurement by measuring three important indexes; Ye Qianyi uses the XGBoost model to mine the attribute data of shopping malls, and extracts the features through the cleaning and visual analysis of the original data, so as to predict the user behavior information and the sales of shopping malls. Although the above algorithms have used the existing algorithms to optimize the feature ranking algorithm, the feature ranking algorithm model still has the problems of low accuracy or too complex model.

To solve the above problems, there is an urgent need for an effective data analysis algorithm to extract the key information of massive data. The first mock exam shows that compared with logistic regression and decision tree, the integration of different models with a certain strategy has higher accuracy and better stability. In recent years, XGBoost algorithm has been widely used in the field of data analysis and provides a method to calculate feature ranking, but the model is complex and it is difficult to adjust parameters. MIC algorithm is often used in feature selection of machine learning. The algorithm is simple and its effect is obviously better than other similar algorithms. Therefore, this paper combines XGBoost and MIC model to design an algorithm to extract the key information of data. The results of the first mock exam of the anti breast cancer candidate drug dataset show that the algorithm can effectively reduce the error value caused by a single model and improve the generalization ability of the model.

II. XGBOOST-MIC COMBINATION MODEL

XGBoost model calculates the importance score of each attribute by gradient lifting algorithm, which has the advantages of preventing over-fitting, simple models and strong flexibility. But it is not suitable for processing high-dimensional feature data, and there are many parameters, so it is relatively difficult to adjust parameters. The MIC algorithm model just needs a large data samples, which can calculate the high-dimensional features well. MIC algorithm can cover all functional relations evenly when the sample size is large enough, rather than limited to some functional expressions.In order to combine the advantages of the two models, a new algorithm XGBoost-MIC combination model, is obtained by combining and weighting XGBoost and MIC models.

A. XGBoost model

XGBoost, the full name of eXtreme Gradient Boosting, is an optimized distributed gradient boosting library, which is efficient, flexible and portable. XGBoost is a tool for large-scale parallel Boosting tree. XGBoost is an optimization of boosting algorithm, which integrates weak classifier into a strong classifier. XGBoost algorithm generates a new tree through continuous iteration to fit the residual of the previous tree.

With the increase of iteration times, the accuracy continues to improve. It is the fastest and the best open source Boosting tree toolkit at present, which is more than 10 times faster than common toolkits. In terms of data science, XGBoost is a necessary weapon for major data science competitions, and a large number of contestants like to choose XGBoost for data mining competitions. In terms of industrial large-scale data, the distributed version of XGBoost is widely portable. And it is supported in various distributed environments such as Kubernetes, Hadoop, SGE, MPI, Dask, etc, which makes it a good solution to the problem of industrial large-scale data.

The definition function of XGBoost is shown in formula (1).

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(X_i) \tag{1}$$

The goal of formula (1) is to learn such K tree models f(x). In order to learn the model f(x), the objective function shown in formula (2) is defined.

$$L(\phi) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega(f_k) \\ where \quad \Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\lambda\|\omega\|^2 \tag{2}$$

Among them, k is the total number of trees, f_k is the model of the kth tree, $\hat{y}_i$ represents the predicted value of the model and $\mathrm{y_i}$ represents the category label of the ith sample, T represents the number of leaf nodes of each tree, and ω represents the set of scores of leaf nodes of each tree.

The tree model of XGBoost and the method of calculating feature importance are introduced below.

1) Tree model

Decision tree is a basic classification and regression method. It is a decision analysis method to judge the project risk and feasibility by constructing a decision tree based on the known probability of various situations. As a supervised learning model, decision tree does not need any prior assumptions about data, so as to quickly find decision rules according to the characteristics of data. XGBoost adopted the gradient lifting algorithm to continuously reduce the loss caused by the last decision tree and generate new models to ensure the reliability of the final decision tree.

In the process of building a tree model with a given data set, greedy algorithm is used to select a feature segmentation point in each layer as a leaf node. If the gain value of the whole tree is the largest after segmentation, it means that the more times the feature is segmented, the better the effect of the whole tree is gained. That means the feature is more important. The weight of leaf nodes in the process of feature segmentation is recorded as $w(g_i, h_i)$, which g_i and h_i are expressed by formula (3) and formula (4) respectively.

$$g_i = \delta_{\hat{\mathrm{y}}^{(\mathrm{t\text{-}1})}} l(y_i, \hat{\mathrm{y}}^{(t-1)}) \tag{3}$$

$$h_i = \delta^2_{\hat{\mathrm{y}}^{(\mathrm{t\text{-}1})}} l(y_i, \hat{\mathrm{y}}^{(\mathrm{t\text{-}1})}) \tag{4}$$

l indicates the gap between the target value y_i and the predicted value $\hat{y}$. Combining the weights of all leaf nodes, in order to maximize the gain value of the segmented tree, each feature can be used as the gain of the separation point, and its gain is the difference between the total weight after segmentation and the total weight of the leaf nodes before segmentation.

2) Feature importance model

Importance is a measure to evaluate the importance of each feature in the feature sets to which it belongs. XGBoost uses three attributes to measure the importance of features, including Freq, Gain and Cover. Here, Freq is the percentage of times that a specific feature occurs in the model tree, Gain is the relative contribution value obtained by calculating the contribution of features in the model to each tree, and Cover is the coverage index of all functions related to a certain function.

The calculation formulas of the above three importance metrics are shown below (5).

$$\begin{aligned} &F\mathrm{req} = \left|X\right| \\ &Gain = \frac{\sum Gain_x}{FScore} \\ &\mathrm{Cover} = \frac{\sum Cover_x}{FScore} \end{aligned} \tag{5}$$

X is a set of feature classification into leaf nodes, Gain is the gain value of each leaf node in X when it is segmented, and Cover is the number of samples falling on each node in X .

In order to get the best model, we construct L = f(θ), and get θ by "gradient descent". With the support of the ascension tree, what we are looking for θ is a parameter for all trees(tree structure, scores of leaf nodes). In order to optimize the parameters and reduce the amount of calculation, we regard a tree as θ ,and build a relationship between the loss function and a certain tree, and then take the derivative of loss for θ , and the process of finding the tree is the solution process of XGBoost.

In this paper, XGBoost model needs to determine three parameters: general parameters, auxiliary parameters and task parameters. General parameters are used to control the function macroscopically; The auxiliary function controls the iterative model, select tree model or linear model; Task parameters control the performance of learning tasks and learning objectives.

Among the above three parameters, the auxiliary parameters have the greatest impact on the algorithm performance, and the maximum height of the tree will affect the final result. Therefore, first tune the maximum height of the tree. In the tuning process, first give other parameters an initial value, and set the important parameter lines to common typical values or default values. Compare the changes of test data by changing the height of the tree, So as to get better parameter selection. After determining the maximum height of the tree, the best combination of other parameters is obtained by traversal method.

Figure 1 shows the process of building decision tree and optimizing parameters. Figure 2 shows the XGBoost algorithm flow.

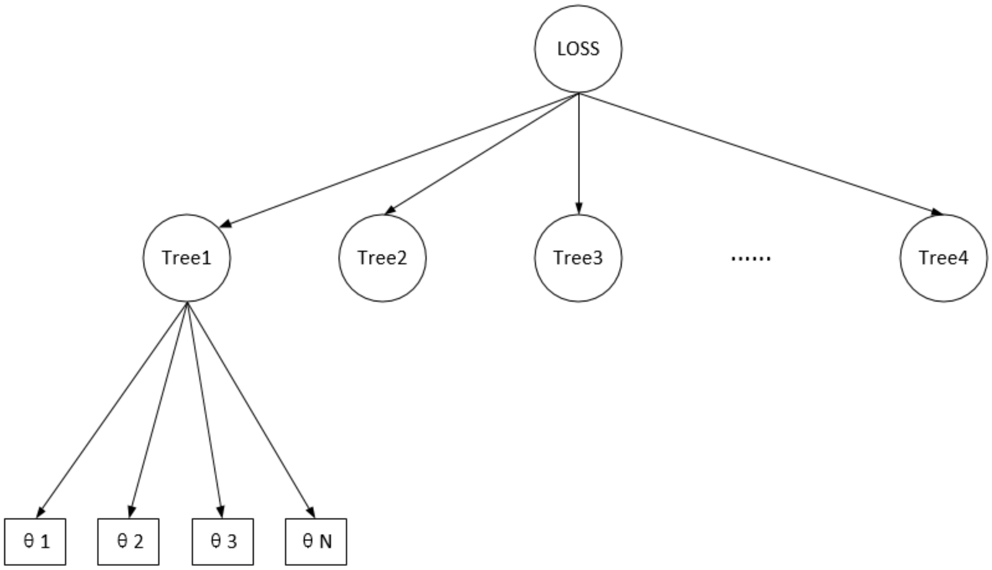


Figure 1. Building decision tree

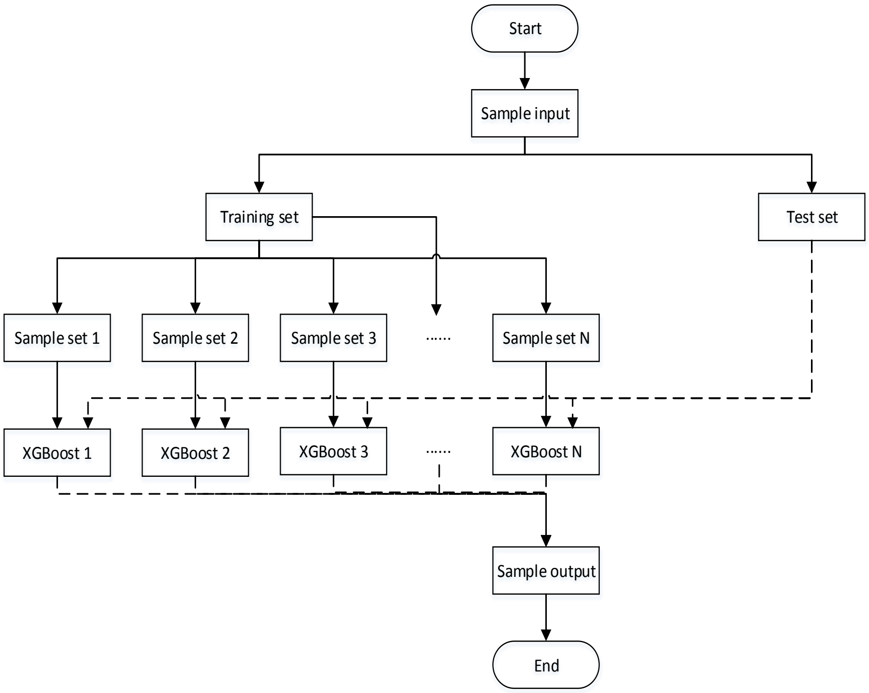


Figure 2. XGBoost algorithm flow

B. MIC correlation coefficient model

MIC (Maximum Mutual Information Coefficient) can be used to measure the degree of correlation between two variables X and Y, which is often used for feature selection of machine learning. Compared with Mutual Information(MI), MIC has a higher accuracy and is an excellent data correlation calculation method.

According to the nature of MIC, MIC is universal, fair and symmetrical. The so-called universality means that when the sample size contains most of the information of the total sample, it can capture all kinds of associations, but not limited to specific function types (such as linear function, exponential function or periodic function), or it can cover all functional relationships in a balanced way. The complex relationship between general variables can be modeled not only by a single function, but also by superposition functions. The so-called fairness means that when the sample size is large enough, similar coefficients can be given for the correlation of different types of single noise with similar degree. Symmetry means that the amount of information obtained from different footholds is the same.

The basic principle of MIC will make use of the concept of mutual information. Mutual information is a measure of the degree of interdependence between random variables, which can be regarded as the information of another variable contained in one random variable. Mutual information can be defined by formula (6). P(x,y) is the joint probability between variables x and y.

$$
\begin{aligned}
I(x,y) &= \int p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}\,dxdy \\
I[x;y] \approx I[X;Y] &= \sum\nolimits_{XY} \log_2 \frac{p(X,Y)}{p(X)p(Y)}
\end{aligned}
\tag{6}
$$

The following formula (7) of MIC model is given as follows.

$$
\begin{aligned}
MIC(x;y) &= \max_{a*b<B} \frac{I(x;y)}{\log_2 \min(a,b)} \\
MIC[x;y] &= \max_{|X||Y|<B} \frac{I[X;Y]}{\log_2(\min(|X|,|Y|))}
\end{aligned}
\tag{7}
$$

MIC model calculates the relationship between two attributes in turn, discretizes them in two-dimensional space, and uses scatter plot to represent them. The model divides the current two-dimensional space into a certain number of intervals in the X and Y directions, and then checks the situation that the current scatters fallen into each grid, thus solving the problem that the joint probability in mutual information is hard to find.

The calculation of MIC value is generally divided into the following three steps: first normalize the data volume through different meshing methods, then calculate the maximum mutual information value, then normalize the maximum mutual information value, and finally select the maximum mutual information value under different scales as the mic value.

C. Combination model

In XGBoost algorithm, the steps of constructing decision tree are as follows:

- Traverse all feature nodes from the root node. For a certain feature, according to the sample value, sort and then determine the segmentation point with the best gain.
- Select the feature with the highest gain from all the selected segmentation points for segmentation.
- Divide to the maximum depth and build the next tree.
- Integrate all trees to complete the model construction.
- The feature importance measure is calculated according to the feature importance index.

According to the above algorithm ideas, the results of the two algorithms are weighted and combined by the error reciprocal method which are shown in formula (8).

$$f_t = \alpha_1 f_{1t} + \alpha_2 f_{2t} \ , \ t = 1,2,3,...,n$$
$$\alpha_1 = \frac{\varepsilon_2}{\varepsilon_1 + \varepsilon_3}, \quad \alpha_2 = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2} \qquad (8)$$

Respectively, ε_1 and ε_2 is the error between the calculated result and the true value by XGBoost and MIC algorithms. It can be seen from formula (8) that this method all error, thus reducing the error of the whole combination model and obtaining the predicted value closer to the real situation, thus improving the overall prediction accuracy. The error reciprocal method can not only give greater weight to the method with higher prediction accuracy, but also ensure that the error reciprocal method can give greater weight to the method with smaller absolute error value (i.e. the method with high prediction accuracy) at any time. If the number of positive and negative errors of each single prediction error is equal, the error reciprocal method can effectively reduce the combined prediction error and achieve better combination effect.

III. EXPERIMENTAL PROCESS AND RESULT ANALYSIS

This experiment runs under Linux Ubuntu16.04 system, and the framework version adopted is XGBoost1.5.0. In terms of hardware environment, the CPU used in the experiment is Intel(R) XEON W-2133 and the GPU is NVIDIA TITAN XP 12G.

The experimental data comes from the effect of 729 molecular descriptors of 1974 compounds provided by China Postgraduate Mathematical Modeling Competition in 2021 on inhibiting the activity of breast cancer cells, including the biological activity value of compounds on ERα and pIC50 obtained by converting IC50 value. In this data set, the biological activities of various compounds are constant and do not react with each other.

The flow of this experiment is shown in Figure 3. Firstly, preprocess the source data. Secondly, the processed data are calculated by XGBoost algorithm and MIC algorithm, and then weighted and combined by the reciprocal error method to obtain a combination model (namely XGBoost-MIC combination model). Finally, the algorithm results of XGBoost-MIC combination model are compared with those of single model (XGBoost or MIC model), and it is decided whether to adjust the parameters. If the results of the combination model are better than those of the single model, the experimental results will be analyzed. Otherwise, adjust the parameters and retrain the model.

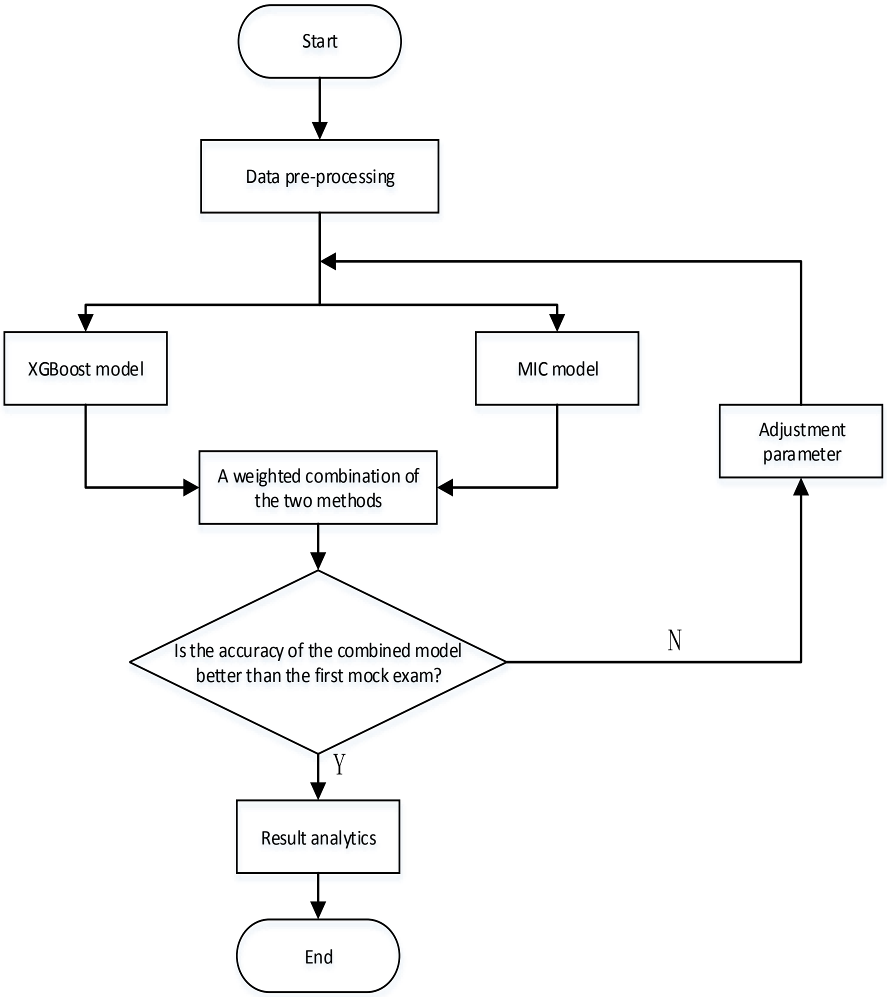

Start
Data pre-processing
XGBoost model
MIC model
Adjustment parameter
A weighted combination of the two methods
Is the accuracy of the combined model better than the first mock exam?
N
Y
Result analytics
End

Figure 3. Experimental flow chart

A. Data preprocessing

In this data set, there are a certain number of unique attributes and empty attributes, so before using this data set, data processing should be carried out first.

For the unique attributes in the data set which can’t describe the distribution law of the sample itself, we can delete these attributes directly.

In order to avoid dealing with high-dimensional data and reduce the difficulty of learning tasks, feature coding is needed. The feature coding method adopted in this paper is One-Hot Encoding. Only One-Hot Encoding uses N-bit status registers to code N possible values, and each state is represented by an independent register, and only one of them is valid at any time.

For data with different magnitudes in the data set, the magnitude difference may greatly affect the calculation results. In order to reduce its effect on the results, the data must be standardized. In this paper, the data are normalized. Its conversion function is shown in formula (9).

$$x = \frac{x - \min}{\max - \min} \tag{9}$$

Finally, we divide the data set into two parts: the training set (80%) and the verification set (20%). The verification set is used to record the accuracy of the algorithm to find out the best model parameters.

B. Experimental results and analysis

As shown in Figure 3, the XGBoost and MIC models are realized respectively. The two models are optimized by grid search method to get the optimal parameters, and then the combination model is calculated by the reciprocal error method. The score and ranking of feature importance calculated by this model are shown in Figure 4.

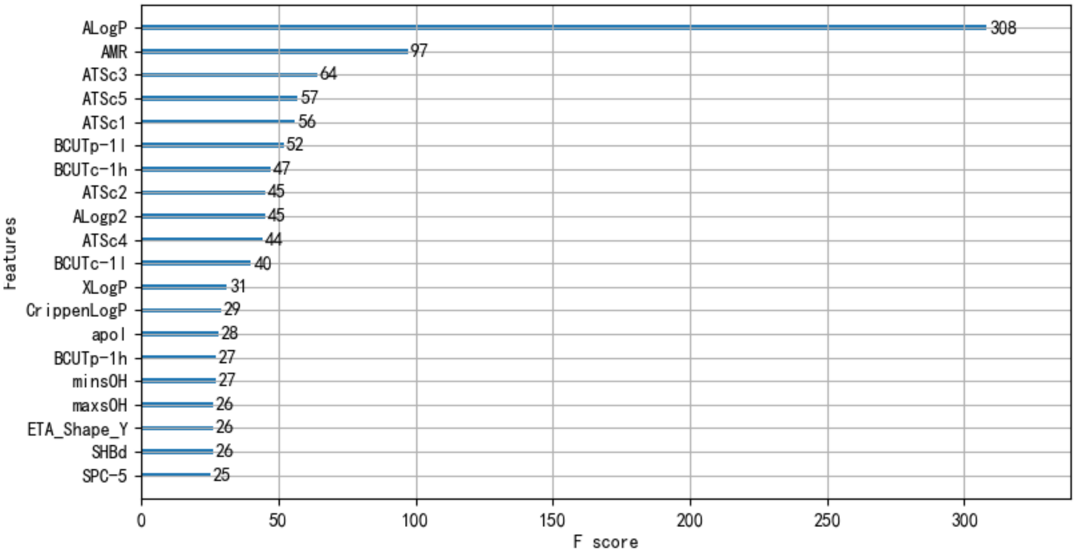


Figure 4. Order of feature importance

In this paper, R^2 (the decisive coefficient of the relationship between a random variable and multiple random variables) is selected to reflect the regression model, and is used as the main evaluation index of the prediction performance of each model. Root mean square error (RMSE) is selected as the auxiliary evaluation index. The calculation of R^2 and RMSE is shown in formula (10) and (11) respectively.

$$\mathrm{R}^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i \left(y_i - \bar{y} \right)^2} \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{m}} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2} \tag{11}$$

y_i is the predicted value of XGBoost-MIC combination model for the I-th data set, $\hat{\mathrm{y}}_{\mathrm{i}}$ is the actual measured value in the I-th data set, and m is the number of samples.

The XGBoost-MIC combination model was tested with the testing-set data divided in the pre-processing, and the R^2 and RMSE values were compared with those of the single model. The results are shown in Table 1.

TABLE I. COMPARSION OF R^2 AND RMSE VALUES BETWEEN XGBOOST-MIC CONBINED MODEL AND SINGLE MODEL.

Method	R²	RMSE
composite pattern	0.759	8.705
XGBoost model	0.723	9.453
MIC model	0.689	8.994

As can be seen from Table 1, the value of XGBoost-MIC combination model is higher(the closer the R^2 value is to the 1, the closer the predicted value is to the real value), The prediction accuracy of the combination model is nearly 0.02 higher than that of the single model. By analyzing RMSE values, it can be concluded that the combination model has lower error than the single model. In general, the combination model has higher accuracy and less error, which has obvious advantages over the single model.

IV. Conclusion

This paper proposes a feature importance selection algorithm based on the combination of XGBoost and MIC. In the first mock exam, the algorithm uses XGBoost to rank the feature importance. The XGBoost-MIC model combines XGBoost and MIC models by weighted reciprocal method. The method automatically improves the model and gives more weight to the smaller error models, which can correct the larger errors caused by the single model. A combination of the first and the first mock exam models was used to compare the data set of the anti breast cancer candidate drugs. The results show that the first mock exam is better than the single model, and the XGBoost-MIC combination model can effectively improve the accuracy and efficiency of the feature ranking, and has a better generalization ability, which is suitable for dealing with large-scale data.

In the future work, we will continue to improve the performance of the algorithm proposed in this paper on a large data set, continue to explore more suitable parameter optimization methods and model fusion methods, continue to process data sets in other fields, and further obtain more robust models combined with scenarios.

References

[1] Zhang Yu, Zhang Yansong, Chen Hong, Susan Wang. OLAP foreign key join algorithm for MIC coprocessor [J]. Journal of Software, 2017, 28(03):490-501.

[2] Yang Guijun, Xu Xue, Zhao Fuqiang. User score prediction model based on XGBoost algorithm and its application [J]. Data Analysis and Knowledge Discovery, 2019, 3(01):118-126.

[3] Ye Qianyi, Rao Hong, Ji Mingshu. Commercial sales forecast based on Xgboost [J]. Journal of Nanchang University (Science Edition), 2017, 41(03):275-281.

[4] Chen Zhenyu, Liu Jinbo, Jerry Lee, Ji Xiaohui, Li Dapeng, Huang Yunhao, Di Fangchun, Gao Xingyu, Xu Lizhong. Ultra-short term power load forecasting based on LSTM and XGBoost combined model [J]. Power Grid Technology, 2020, 44(02):614-620.

[5] Zhang Chengchang, Zhang Huayu, Luo Jianchang, He Feng. Analysis method of massive electricity consumption data based on cloud computing and improved K-means algorithm [J]. Computer Applications, 2018, 38(01):159-164.

[6] Liu Nian, Liu Yu. Research on visualization technology of massive relational data based on clustering analysis algorithm [J]. Electronic Design Engineering, 2018, 26(10):92-95.

[7] Cheng Xueqi, Jin Xiaolong, Wang Yuanzhuo, Guo Jiafeng, Zhang Tieying, Li Guojie. Overview of Big Data System and Analysis Technology [J]. Journal of Software, 2014, 25(09):1889-1908.

[8] Zhou Yanjun, Wang Shuangcheng, Wang Hui. Research on classifier based on Bayesian network [J]. Journal of Northeast Normal University (Natural Science Edition), 2003(02):21-27.

[9] Xuanxuan Lin. Research on Enterprise Bankruptcy Prediction Method Based on XGBOOST Model [A]. Wuhan Zhicheng Times Cultural Development Co., Ltd. proceedings of 4th international conference on e-education, e-business and information management (EEIM 2021) [c]. Wuhan Zhicheng Times Cultural Development Co., Ltd.: Wuhan Zhicheng Times Cultural Development Co., Ltd., 2021:8.

[10] Shenglong Li,Xiaojing Zhang. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm [J]. Neural Computing and Applications, 2020, 32(prepublish):

[11] Feng Chen, Chen Zhide. Application of xgboost and LSTM weighted combination model in sales forecast [J]. Computer system application, 2019, 28 (10): 226-232. Doi: 10.15888/j.cnki.csa.007091

[12] Shenglong Li,Xiaojing Zhang. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm [J]. Neural Computing and Applications, 2020, 32(prepublish):

[13] Wei Dong,Yimiao Huang,Barry Lehane,Guowei Ma. XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring [J]. Automation in Construction, 2020, 114(C):

[14] Chixiang Wang,Junqi Guo. A data-driven framework for learners’ cognitive load detection using ECG-PPG physiological feature fusion and XGBoost classification [J]. Procedia Computer Science, 2019, 147:

[15] Gao Yifan, Yu Wenzhe, Chao Pingfu, et al. Score prediction and recommendation based on comment analysis [J]. Journal of East China Normal University: Natural Science Edition, 2015 (3): 80-90. (Gao Yifan, Yu Wenzhe, Chao Pingfu, et al. Analyzing reviews for rating prediction and item recommendation [J]. Journal of East China Normal University: Natural Science, 2015 (3): 80-90.)

[16] Li V.,Costantino H.,Rowland J.,Yue L.,Gupta S.. ML3 LASSO (Least Absolute Shrinkage and Selection Operator) and XGBoost (eXtreme Gradient Boosting) Models for Predicting Depression-Related Work Impairment in US Working Adults [J]. Value in Health, 2021, 24(S1).

[17] Li V.,Costantino H.,Rowland J.,Yue L.,Gupta S.. ML3 LASSO (Least Absolute Shrinkage and Selection Operator) and XGBoost (eXtreme Gradient Boosting) Models for Predicting Depression-Related Work Impairment in US Working Adults[J]. Value in Health, 2021, 24(S1).

[18] Deng Xiaoyi, Jin Chun, Han Qingping, et al. Collaborative filtering recommendation model based on situational clustering and user rating [J]. System engineering theory and practice, 2013, 33 (11):2945-2953. (Deng Xiaoyi, Jin Chun, Han Jim C, et al. Improved Collaborative Filtering Model Based on Context Clustering and User Ranking [J]. Systems Engineering −Theory & Practice, 2013, 33(11): 2945-2953.)

[19] Zhang Hongli, Liu Jiying, Yang Sinan, et al. Research on scoring prediction model based on Internet user comments [J]. Data analysis and knowledge discovery, 2017, 1 (8): 48-58

[20] McLachlan P, Munzner T, Koutsofios E, et al. LiveRAC: interactive visual exploration of system management time-series data[C] //Proceeding of the 26th International Conference on Human Factors in Computing Systems. New York: ACM Press, 2008: 1483-1492.

[21] Agryzkov T, Oliver J L, Tortosa L, et al. Analyzing the commercial activities of a street network by ranking their nodes: a case study in Murcia, Spain [J]. International Journal of Geographical Information Science, 2014, 28(3/4): 479-495.

[22] Qiu X, Suganthan P N, Amaratunga G A J. Ensemble incremental learning random vector functional link network for short-term electric load forecasting [J]. Knowledge-Based Systems, 2018, 145(4):182-196.

[23] Wang J, Lou C, Yu R, et al. Research on hot micro-blog forecast based on XGBOOST and random forest [M]. Knowledge Science, Engineering and Management, KSEM 2018, Lecture Notes in Computer Science, Springer.

[24] Li C, Chen Z.Y, Liu J.B, et al. Power load forecasting based on the ,combined model of LSTM and XGBoost [C]//PRAI ' 19:Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence. Wenzhou, China: ACM, 2019: 46-51.

[25] Gómez-Ríos A, Luengo J, Herrera F. A study on the noise label influence in boosting algorithms : Adaboost, GBM and XGBoost [C]//International Conference on Hybrid Artificial Intelligence Systems (HAIS), 2017:268-280.

[26] Yue Yanchun, Huang Tingzhu. Error reciprocal variable weight combination prediction method [J]. Journal of University of Electronic Science and technology, 2007 (S1): 349-351.

[27] Cao y B, Xu J, Liu T Y, et al. Adapting ranking SVM to document retrieval[C] //Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2006: 186-193.

[28] Xu P P, Mei H H, Ren L, et al. ViDX: visual diagnostics of assembly line performance in smart factories [J]. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1):291-300.

[29] Sun Guanglu, song Zhichao, Liu Jinlai, Zhu Suxia, he Yongjun. Feature selection method based on maximum information coefficient and approximate Markov blanket [J]. Journal of automation, 2017, 43 (05): 795-805. Doi: 10.16383/j.aas.2017.c150851.

[30] Zhang Li, Yuan Yuyu, Wang Zong. Fcbf feature selection algorithm based on maximum correlation information coefficient [J]. Journal of Beijing University of Posts and telecommunications, 2018, 41 (04): 86-90. Doi: 10.13190/j.jbupt.2017-229.

Research on Hierarchical Multi-core Scheduling Algorithm Based on Task Replication

Yin Haijing
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 1391275494@qq.com

Huang Shujuan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 349242386@qq.com

Wang Jianguo
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 2269261628@qq.com

***Abstract*—The rapid development of multi-core systems makes task scheduling in multi-core systems a new research topic. While tasks are running in parallel, how to improve the efficiency of the system and maintain the load balance of the system is the focus of research in the new era. Aiming at the problem that the multi-core scheduling algorithm based on task duplication does not consider the load balance of each CPU, which leads to the problem of reduced CPU utilization. This paper combines a hierarchical idea on the basis of task replication, and proposes a new hierarchical multi-core scheduling algorithm TDLS algorithm based on task replication. This algorithm is based on the idea of hierarchical scheduling. According to the fact that there is no dependency relationship between tasks at the same layer after layering, the task scheduling sequence is adjusted to reduce the waste on the core, shorten the waste between cores caused by communication time, and reduce the number of processors. , Thereby greatly improving the CPU utilization rate, using the least time and the least number of cores to complete scheduling, making the load of multi-core scheduling more balanced. Experiments show that under the same experimental conditions, compared with the traditional multi-core scheduling algorithm based on task replication, the improved algorithm TDLS reduces the number of processor cores, and also shortens the scheduling length of the total task. Its performance is better than the traditional multi-core scheduling algorithm based on task replication.**

Keywords-Load Balancing; Task Scheduling; Task Duplication; Hierarchical Scheduling

I. INTRODUCTION

Multi-core processor technology mainly integrates two or more processor cores on a single chip to enhance computing performance. Multi-core processors improve system performance by distributing load on multiple CPU

DOI: 10.21307/ijanmc-2021-038

cores, and relying on high-speed on-chip interconnection and high-bandwidth pipelines of memory and input/output (I/O). Under the same conditions, multi-core processors can bring more performance and productivity advantages than current single-core processors. Therefore, the research of scheduling algorithms under multi-core platforms is also a future development trend.

Multi-core processor task scheduling refers to how to allocate multiple tasks to multiple cores for parallel execution through a scheduling algorithm, so as to minimize the total time for task completion. Multi-core task scheduling has long been proved to be an NP problem [1], and it is difficult to find the optimal solution in polynomial time. The most common task scheduling algorithm is based on heuristic scheduling algorithm. Heuristic scheduling algorithms mainly include table scheduling algorithm based on critical path [2-5], task duplication algorithm [6-8], processor allocation algorithm based on task duplication [9], improved multi-core scheduling based on task duplication Algorithm [10]，clustering algorithm [11-13] and so on.

Since the communication overhead between tasks on the same processor can be ignored, scheduling based on task duplication is an effective strategy for reducing communication overhead. The characteristic of the task copy method is to reduce the communication time between processors by copying the predecessor tasks that have a communication relationship, thereby reducing the execution time of the system as a whole.

When using reasonable and effective duplication rules and strategies, scheduling algorithms based on task duplication have been proven to have better scheduling effects than other scheduling algorithms. However, the scheduling algorithm does not consider the factor of load balancing, and in the DAG graph, there is no dependency between nodes in the same layer. According to the adjustment of the scheduling sequence between nodes in the same layer, the idle time is reduced and the CPU is increased. Utilization, while coordinating the load in each CPU to make it more balanced. Therefore, this paper proposes a hierarchical scheduling algorithm based on task duplication to solve the shortcomings of unbalanced load of traditional scheduling algorithms based on task duplication.

II. TASK SCHEDULING MODEL

The task scheduling problem is a kind of combinatorial optimization problem in mathematics, that is, an abstract task model of a computer application is established, and then based on the constraints of the task model, through a reasonable scheduling strategy, a scheduling sequence is generated and the tasks are assigned to the processing cores for calculations. Get the least total task execution time and maximize the parallel execution advantages of multi-core systems.

The task scheduling model is mainly divided into two aspects: system model and task model. The system model is a mathematical abstraction of information such as the topological structure and computing capabilities of a multi-core system, and the task model is a mathematical abstraction of computer application programs. It mainly includes information such as the constraint relationship between tasks and the characteristics of the task itself. The following are two parts Detailed discussion.

A. System model

The system model is an abstraction of the actual computing system. The actual computing system in this article is a multi-core system, that is, a system composed of multiple processing cores.

The system is generally expressed as $\mathrm{P}=\{\mathrm{p}_1,\mathrm{p}_2,\cdots,\mathrm{p}_i,\cdots,\mathrm{p}_n\}$.

Among them, P represents a collection of processing cores in a multi-core system, which p*i* represents the i processing core, and n represents that the system contains a total of n cores.

B. Mission model

The relationship between multi-core tasks is generally represented by DAG (Directed Acyclic Graph), and when there is a dependency between tasks, a weighted DAG graph is used (as shown in Figure 1)

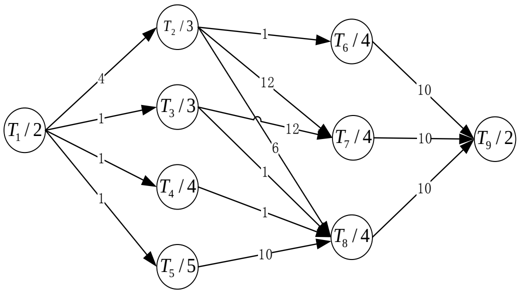


Figure 1. DAG diagram

Its mathematical description is:

$$G=\{T,E,t,c\}. \tag{1}$$

Among them, the formula $T=\{T_1,T_2,\cdots,T_i,\cdots,T_n\}$ represents the set of nodes in the graph, which is the first task; represents the set of nodes in the graph, which is the first task;

$E=\{E_{ij}\}$ Represents the set of directed edges that E_{ij} is a communication relationship between task T_i and task T_j, otherwise, they cannot communicate directly. $t=\{t_1,t_2,\cdots,t_i,\cdots,t_n\}$. This Set represents the set of node weights in the graph，in other words，t_i is execution time of the task *i*. Meanwhile, The set,$\{c=c_{ij}\}$,is the set of weights of directed edges that c_{ij} Indicates the communication time between task T_i and task T_j.

Since the communication time of tasks between different cores is much longer than the communication time between the same cores, when two tasks are on the same processor, the communication time is ignored, that is $c_{ij}=0$ in two related tasks on the same processor.

Definition:

- The earliest start time $T^i{}_{begin}$ of task *i*: it represents the smaller value between the predecessor time of task execution and the maximum associated predecessor time in the task predecessor set. which is:

$$T^i{}_{begin}=\min\{T^i{}_{exe\_pre},\max\{T^i{}_{re\_pre}\}\}. \tag{2}$$

- The completion time $T^i{}_{end}$ of task *i*: the time when the task is executed on the processing core is equal to the start time plus the task execution time, which means the time it takes to complete the task. which is:

$$T^i{}_{end}=T^i{}_{begin}+t_i. \tag{3}$$

- Associated predecessor *j* of task *i*: The set of tasks that must be completed before the task *i* is executed. That is, the task set on which the execution of the task depends. For example, the task set T_i, Tj in Figure 1 is the associated predecessor of task T_7. The associated predecessor time $T^i{}_{re\_pre}$ of task *i* is: if the associated predecessor task *j* and task *i* of the task are on the same core, the associated predecessor time is the completion time of the associated predecessor task *j*; if not on the same core, The associated predecessor The time is the completion time of the associated predecessor task *j* and the maximum value of the sum of communication values between task *i* and its associated predecessor task *j*.

$$T^{i}{}_{re\_pre} = \begin{cases} T^{j}{}_{end}, & i, j \text{ in the same core} \\ T^{j}{}_{end} + c_{ij}, & i, j \text{ in a different core} \end{cases}. \quad (4)$$

Where task *j* is a predecessor task of task *i*.

- The running result of a certain processor $p^{i}{}_{end}$ indicates the total time spent by the processor after all tasks on the processor are scheduled.
- The running result of all processors P_{end}: it represents the final task scheduling result, namely: the total time for all tasks to complete.
- The successor task *next* of task *i*: the task related to task *i*, and it must be ensured that task *i* has been executed before it is executed.
- Execution predecessor *k* of task *i*: On the same core, task *k* to be executed before task execution is the execution predecessor of task *i*. And the execution predecessor time $T^{i}{}_{re\_pre}$ of task *i* is: the completion time of the execution predecessor of task *k*, that is:

$$T^{i}{}_{re\_pre} = T^{k}{}_{end}. \quad (5)$$

- The idle time $T^{i}{}_{rest}$ of task *i*: represents the wasted time slice on the same core when the task is executed. which is:

$$T^{i}{}_{rest} = T^{i}{}_{begin} + T^{i}{}_{exe\_pre}. \quad (6)$$

- The communication start time $T^{next}{}_{begin\_comm}$ between task *i* and the successor task *next*: It mainly represents the communication relationship between task *i* and its successor task *next* that are not on the same processor. That is: after the execution of task *i* is over, after the communication time between processor cores, the start time of the successor task *next*.

C. Basic constraints of task scheduling

For a certain task, when it meets the following two necessary conditions, it can be executed on a specific processing core.

On one hand, All the predecessor tasks of the same processing core with which it is dependent have been executed, and the communication between the predecessor tasks that are not on the same processing core and this task has also been completed;

On the other hand, the time period occupied between the task start time and the end time does not conflict with other tasks on the processing core where the task is located.

- For task *i*, its earliest start execution time must not be less than the execution completion time of all predecessor tasks, which is:

$$T^{i}{}_{begin} \geq \max\{T^{i}{}_{re\_pre}\}. \quad (7)$$

- The start time of the communication between the task and the successor task must be after the end time of the task, because only the task execution is completed before the relevant data can be provided to the successor task, which is:

$$T^{i}{}_{begin\_comm} \geq T^{i}{}_{end}. \quad (8)$$

- For a task graph, the final task scheduling result depends on the task that finishes executing at the latest. That is, the time used by the processor with the longest scheduling length among all processors. which is:

$$\mathrm{P}_{\mathrm{end}} = \max\{ p^{i}{}_{end} \}. \qquad (9)$$

III. TASK DUPLICATION SCHEDULING ALGORITHM

The idea of task duplication scheduling algorithm is to generate copies of specific tasks through duplication. These copies will be allocated to the processing core according to a certain strategy. When the subsequent tasks of the copy are allocated to the same processing core, they can be offset. Consumption of communication between tasks to save time.

Task duplication can be divided into single-task duplication and multi-task duplication according to the number of duplication tasks. Single-task duplication copies and allocates tasks that restrict the start time of the current task to the processing core; multi-task duplication copies and allocates multiple predecessor tasks of the task to the processing core.

Taking Figure 1 as an example, the successor tasks of task T_1 are T_2, T_3, T_4, and T_5. Therefore, when T_1 and T_2, T_3, T_4, and T_5 are executed on different processing cores, there will be inter-core communication, in order to save inter-core During the communication time, a task duplication strategy is adopted to replicate T_1, and all three copies are allocated to the processing cores where T_2, T_3, T_4, and T_5 are located. Through this task duplication method, the task scheduling result is finally optimized.

As shown in Figure 2, if the T_1 task is not copied, and assuming that the tasks T_2 and T_1 are not on the same core, the earliest start time of the task T_2 is equal to the time W_1 waiting for the completion of the predecessor task T_1 and the task that is not on the same processor The sum of the communication time $C_{1,2}$ between T_1 and task T_2, that is, the earliest start time of T_2 is 6, and the scheduling result of task T_2 is 9; and if the task duplication strategy is adopted for scheduling, a copy is made on the processor where T_2 is located The copy of T_1, at this time, because the communication time is reduced by 4, the start time of T_2 becomes 2, and the scheduling result of task T_2 is reduced to 5 (as shown in Figure 3). At this time, the optimization effect is significant and the scheduling efficiency is greatly improved.

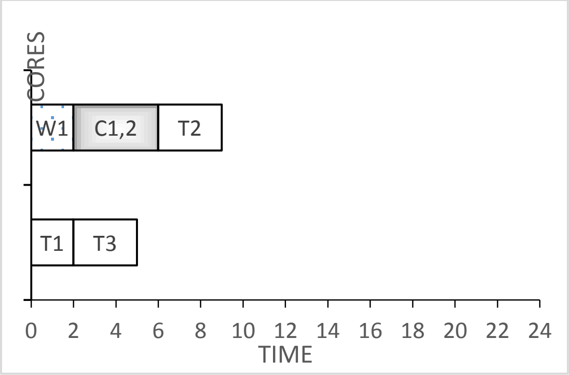


Figure 2. T1 does not use task duplication strategy T2 scheduling result diagram

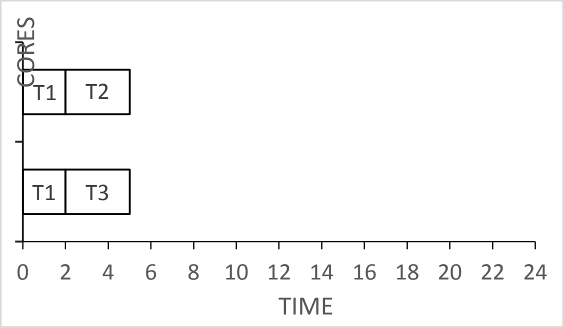


Figure 3. T1 adopts task duplication strategy T2 scheduling result graph

IV. TDLS SCHEDULING ALGORITHM

In the DAG diagram, tasks can be classified by layer. The tasks in the same layer are independent, and the tasks in the same layer are not dependent on each other, that is, if there is no priority difference, the tasks in the same layer, there is no difference in the execution of tasks. The TDLS algorithm mainly relies on hierarchical scheduling without dependencies between tasks on the same layer. By adjusting the scheduling sequence of tasks on the same layer, tasks with a smaller initial start time can be scheduled when the time slice comes, reducing the number of cores. At the same time, considering that the communication value between tasks with dependencies between different layers is too large, it will also affect the completion time of the task, so the predecessor tasks that have a greater impact on the task are adjusted to reduce the communication time between tasks, Thereby increasing the CPU utilization.

A. Steps of hierarchical multi-core scheduling algorithm based on task replication

Step 1: Calculate the in-degree of all tasks in the DAG graph;

Step 2: Determine whether the in-degree of all tasks is 0, and put all tasks with in-degree of 0 into one layer, that is, the k layer is obtained;

Step 3: Remove the tasks that have been layered in the DAG graph and their related edges to get a DAG graph, and make $k = k + 1$, repeat steps 1~3, until the task in the DAG graph is empty, that is The DAG graph completes the layering operation.

B. The Steps of TDLS Algorithm

Step 1: According to the in-degree of the tasks in the DAG, use the hierarchical algorithm to perform hierarchical operations on all tasks;

Step 2: According to the layering result, the scheduling sequence is initially obtained; because the tasks in the same layer do not have mutual dependence, the tasks of the same layer can be scheduled according to the earliest start time T^{i}_{begin} of the task in the order of scheduling from small to large to reduce idle time , Improve CPU utilization.

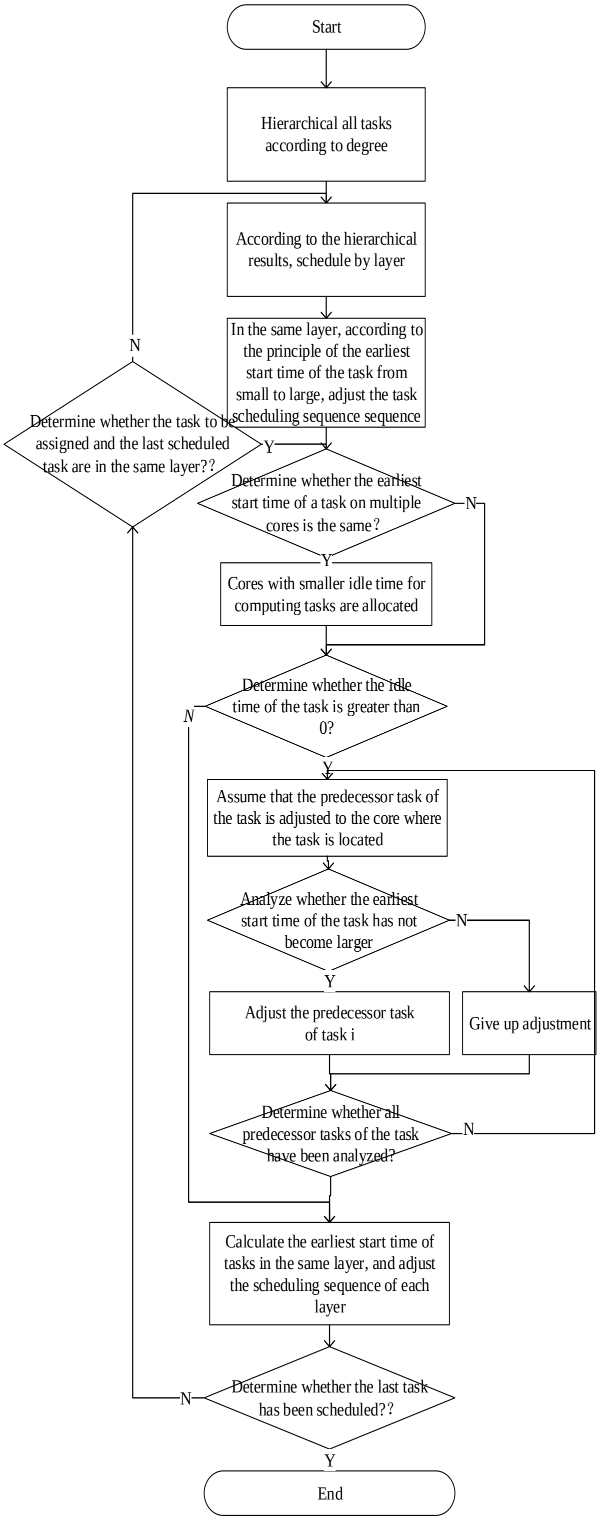

Start
Hierarchical all tasks according to degree
According to the hierarchical results, schedule by layer
In the same layer, according to the principle of the earliest start time of the task from small to large, adjust the task scheduling sequence sequence
N
Determine whether the task to be assigned and the last scheduled task are in the same layer??
Y
Determine whether the earliest start time of a task on multiple cores is the same?
N
Y
Cores with smaller idle time for computing tasks are allocated
Determine whether the idle time of the task is greater than 0?
N
Y
Assume that the predecessor task of the task is adjusted to the core where the task is located
Analyze whether the earliest start time of the task has not become larger
N
Y
Adjust the predecessor task of task i
Give up adjustment
Determine whether all predecessor tasks of the task have been analyzed?
N
Calculate the earliest start time of tasks in the same layer, and adjust the scheduling sequence of each layer
N
Determine whether the last task has been scheduled??
Y
End

Figure 4. TDLS algorithm flow chart

Step 3: Schedule each task in turn according to the adjusted task scheduling sequence sequence. The polling method assigns the task T_i to each core in turn, calculates and compares the earliest start time T^i_{begin} of the task on each core, so as to find out which core takes the least time, that is, allocate T_i to the core.

If the task's earliest start time T^i_{begin} is not single, consider the core with the smaller idle time T^i_{rest} for priority allocation, and when the idle time T^i_{rest} is greater than 0, it means that the communication time between a predecessor task $T^i_{re\_pre}$ of the task and the task is relatively large. When they are not on a core, the earliest start time T^i_{begin} of the task is relatively large, so the precursor task $T^i_{re\_pre}$ that affects the start time of the task is adjusted to the core where the task is located for analysis and comparison. If the earliest start time of the task is not If it changes or becomes smaller, adjust the predecessor task $T^i_{re\_pre}$ of the task to the core where the task is located; otherwise, give up the adjustment. The final scheduling sequence is the optimal scheduling sequence (the algorithm flow chart is shown in Figure 4).

C. Analysis of the time complexity of TDLS algorithm

When selecting an algorithm for a task scheduling problem, it is usually necessary to evaluate the time complexity of the algorithm. Time complexity refers to the increasing trend of the execution time required by the algorithm when the scale of the problem expands. Generally, O is used to represent the time complexity. Among them, O(1) means that the time complexity of the algorithm is constant, that is, no matter how much the problem scale increases, the time consumed by the algorithm remains the same; O(n^2) means that the time complexity of the algorithm is square, that is, when the problem scale When the problem is enlarged by 2 times, the time required for the algorithm to solve the problem is 4 times; O(2^n) means that the time complexity of the algorithm is exponential of 2, and once the scale of the problem increases, the time consumed by the algorithm will be exponential increase.

Assume that the DAG task graph has n nodes. By traversing the task graph to schedule each node, the time complexity required for each node is O(n). For each node of each layer, it is necessary to perform simulation scheduling before confirming the scheduling to determine whether the scheduling reduces idle time, reduces CPU waste, and improves CPU utilization. When planning to schedule, the scheduling sequence of each layer needs to be adjusted. At this time, assuming that there are m nodes in a certain layer, the time complexity required at this time is O(m), so the TDLS algorithm is extremely In this case, the time complexity of the algorithm is not greater than O(n^2).

V. Experimental results and analysis

This chapter will use the TDLS algorithm and the traditional task replication-based scheduling algorithm (CPTD and TDMC algorithm) to conduct experiments in the same environment to schedule the specific DAG graph (Figure 1), and use TDLS, CPTD and TDMC to perform the experiments respectively. The task scheduling distribution diagram of the three algorithms available for scheduling is shown in Figures 5~7, and the two performance indicators and the algorithm time complexity of the number of processors used and the CPU utilization rate are compared. Each of the three algorithms is the comparison of performance evaluation parameters is shown in Table 1.

A. Performance evaluation parameters

The performance of the task scheduling algorithm can be evaluated with the following performance evaluation parameters:

- The number of processors used to complete all tasks. After all tasks in the task graph are scheduled, the fewer processors are used, the more resources can be saved.
- Scheduling length. After all tasks in the task graph are scheduled, the length of time used, the smaller the scheduling length, the better the explanation of resources, and the better the algorithm.

B. Task diagram example analysis

For a specific task graph example (Figure 1), call TDLS, CPTD, and the (TDMC algorithm for short) in Literature 14 (referred to as TDMC algorithm [14]) respectively based on task duplication scheduling algorithm, draw a scheduling diagram, and compare these 3 algorithms Various performance indicators.

The DAG graph shown in Fig. 1 has 4 layers, including 9 tasks $T_1 \sim T_9$. Among them, each circle in the figure represents a task node, and the nodes in the node represent the task and the time required to execute the task. The line segment with arrows in the figure represents the communication dependency between tasks. Where the start position of the arrow line segment is the predecessor task node, the end position is the successor task node, and the number on the straight line represents the communication time between tasks (because the execution time of two tasks on the same CPU is much shorter than that of different CPU The execution time of the two tasks between the two, therefore, when two dependent tasks are on the same CPU, the communication time can be ignored).

TABLE I. COMPARISON OF SCHEDULING RESULTS OF TDLS, CPTD, AND TDMC ALGORITHMS

Algorithms	Scheduling Length	Number of Processors	Algorithm Time Complexity
TDLS	23	2	O(n^2)
CPTD	23	3	O(n^2)
TDMC	24	5	O(n^2)

Use TDLS, CPTD and TDMC three algorithms to schedule the DAG graph respectively, and the task scheduling and distribution diagram corresponding to the three algorithms can be obtained, as shown in Figure 5~7, and the comparison of each performance evaluation parameter under the three algorithms the situation is shown in Table 1.

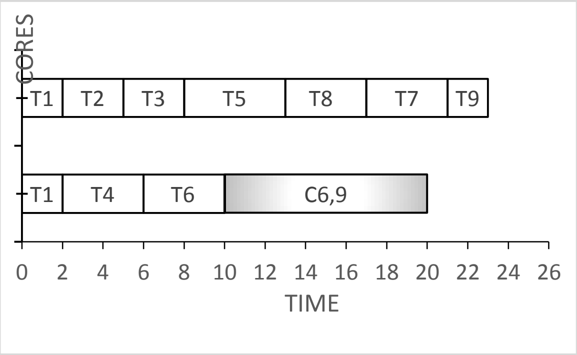


Figure 5. TDLS algorithm scheduling diagram

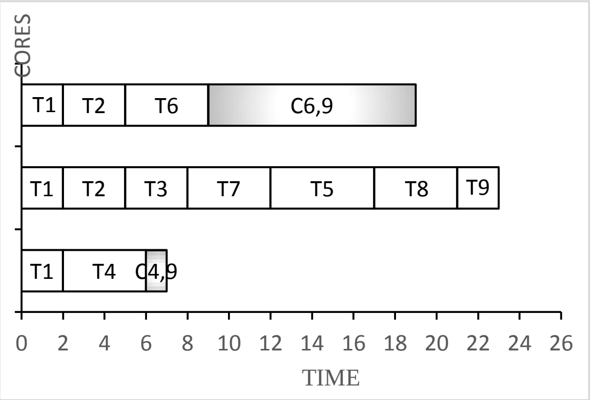


Figure 6. CPTD algorithm scheduling diagram

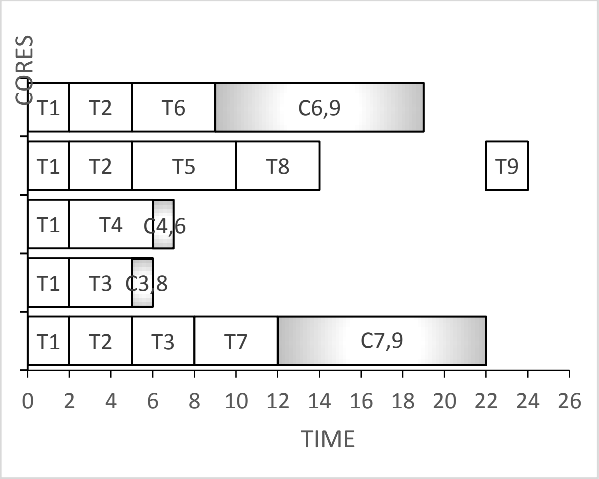


Figure 7. TDMC algorithm scheduling diagram

It can be seen from Figures 5 to Figures 7 and Table 1 that the TDMC algorithm is used to schedule the task graph. The total execution time is 24, and the number of processor cores is 5. The main reason why the CPU is not fully utilized is T_7 and T_9. The communication time of the node is too long, resulting in waste between T_9 and T_8, thereby increasing the total length of task scheduling. The CPTD algorithm is used to schedule the task graph. The total execution time is 23, and the number of processor cores is 3. The algorithm first finds the critical path, and then adopts the early completion strategy of the preceding node group to merge the scheduling sequences of nodes T_7 and T_8 , So that the execution time of T_9 can be advanced, and the communication time can be better controlled. The scheduling of T_6 and T_4 is completed through task replication, but the processors where these two tasks are located mainly complete the scheduling of T_6 and T_4, which greatly reduces the CPU Utilization rate makes the CPU load unbalanced.

The total execution time used by the improved TDLS algorithm is 23, which only occupies 2 processor cores. It is mainly used in the scheduling process of the third layer, by adjusting the scheduling sequence, according to the T_7 and T_8 nodes are unrelated For nodes on the same layer, T_7 is adjusted to T_8 before scheduling, which shortens the waste of time caused by too long communication time, improves CPU utilization, and makes the load more balanced than the previous algorithm.

Compared with the TDMC algorithm, the TDLS algorithm on the one hand shortens the total scheduling length of task execution (the scheduling length is reduced from 24 to 23), and on the other hand, it reduces the number of CPU required to complete all tasks (the number of CPU is reduced from 5 to 2), the algorithm reduces the waste of resources on the whole and improves the utilization of CPU; at the same time, compared with the CPTD algorithm, the TDLS algorithm reduces the number of redundant tasks on the one hand (such as: Reduction in the number of tasks T_1 and task T_2 redundancy). On the other hand, it also reduces the number of CPU required to complete the task (the number of CPU is reduced from 3 to 2). It also reduces the waste of time slices between tasks and greatly improves each CPU The utilization rate makes the load more balanced. Through the comparison and analysis of the above two groups of experiments, it can be seen that the TDLS algorithm has better scheduling performance than the CPTD algorithm and the TDMC algorithm.

VI. CONCLUSION

Aiming at the problem of load imbalance in the traditional multi-core scheduling algorithm based on task duplication under certain circumstances, this paper proposes a hierarchical scheduling algorithm based on task replication, TDLS. TDLS is based on no scheduling sequence between tasks in the same layer. The scheduling sequence of tasks in the same layer can shorten the idle time of the CPU, shorten the execution time of the program, and then improve the utilization of multiple CPU. It is proved through comparative experiments that the TDLS algorithm is compared to the other two traditional scheduling algorithms based on task replication. The load is more balanced, and it also has a certain significance for the improvement of the scheduling performance of the multi-core parallel computer system.

The hierarchical task scheduling algorithm based on task replication proposed in this paper overcomes some inherent shortcomings of traditional task replication-based algorithms in the experimental environment of simulated scheduling, and improves the efficiency of task scheduling and CPU utilization. The research of this algorithm is completely based on assumptions and simulated environment. However, the real situation is more complicated and changeable. For multi-core systems, its load balancing, power consumption, and communication congestion between cores need to be considered under actual conditions. Therefore, it is necessary to further expand the scope of research in future research, apply it to real-time systems, and make more comprehensive considerations for the problems in the actual situation.

REFERENCES

[1] Han Yingjie. Research on Multi-core Task Scheduling Based on Comprehensive Scheduling Critical Path [D]. Harbin University of Science and Technology, 2014.

[2] Ren Liangyu, Zhao Chengping, Yan Hua. Multi-core scheduling algorithm based on task duplication and redundancy elimination [J]. Computer Engineering, 2019, 45(05):59-65.

[3] Shi Wei, Zheng Weimin. A balanced dynamic critical path scheduling algorithm based on related task graphs [J].Chinese Journal of Computers, 2001(09):991-997.

[4] Jing-Jang Hwang, Yuan-Chieh Chow, Frank D. Anger,Chung-Yee Lee. Scheduling Precedence Graphs in Systems with Interprocessor Communication Times. [J]. SIAM J. Comput.,1989,18(2):

[5] Wu M Y, Gajski D D. Hypertool: a programming aid for message-passing systems [J]. IEEE Transactions on Parallel and Distributed Systems, 1990, 1(3):330-343.

[6] Liu Y, Jia P, Yang Y. Efficient scheduling of DAG tasks on multi-core processor based parallel systems[C]// Tencon IEEE Region 10 Conference. IEEE, 2016.

[7] S. Darbha and D. P. Agrawal, "Optimal scheduling algorithm for distributed-memory machines," in IEEE Transactions on Parallel and Distributed Systems, vol. 9, no. 1, pp. 87-95, Jan. 1998, doi: 10.1109/71.655248.

[8] An optimal scheduling algorithm based on task duplication [J]. Journal of Systems Engineering and Electronics, 2005(02):445-450.

[9] Harbin. An Algorithm of Processor Pre-Allocation Based on Task Duplication [J]. Chinese Journal of Computers, 2004.

[10] Ye Jia, Zhou Mingzheng. An improved multi-core scheduling algorithm based on task replication［J］. Computer Engineering and Applications, 2015, 51(12): 31-37.

[11] Boeres C, Filho J V, Rebello V. A Cluster-based Strategy for Scheduling Task on Heterogeneous Processors[C]// Symposium on Computer Architecture & High Performance Computing. IEEE, 2004.

[12] Palis M A, Liou J C, Wei D. Task Clustering and Scheduling for Distributed Memory Parallel Architectures. IEEE Transactions on Parallel and Distributed Systems, 7(1):46-55 [J]. IEEE Transactions on Parallel and Distributed Systems, 1996, 7(1):46-55.

[13] Lan Zhou. Research on scheduling algorithms in distributed systems［D］. Chengdu: University of Electronic Science and Technology of China, 2009.

[14] Zhiqiang Xie, Lei Zhao, Yu Xin, Jing Yang. A Scheduling Optimization Algorithm Based on Task Duplication for Multi-core Processor [J]. Energy Procedia,2011,13:

[15] Ahmad Wakar, Alam Bashir. An efficient list scheduling algorithm with task duplication for scientific big data workflow in heterogeneous computing environments [J]. Concurrency and Computation: Practice and Experience, 2020,33(5):

[16] Computing - Supercomputing; Findings in the Area of Supercomputing Reported from Harbin Institute of Technology (Linear and Dynamic Programming Algorithms for Real-time Task Scheduling With Task Duplication) [J]. Computer Weekly News,2019:

[17] Cao Zhebo, Li Qing. Research and design of multi-core processor parallel programming model [J]. Computer Engineering and Design, 2010, 31(13): 2999-3002+3056.

[18] Chen Gang, Guan Nan, Lu Mingsong, Wang Yi. A review of real-time multi-core embedded systems [J]. Journal of Software, 2018, 29(07): 2152-2176.

Uncovering Correlations Between Urban Road Network Centrality and Human Mobility

Yury Halavachou
Department of Railway Transportation Control
Belarusian State University of Transport
34, Kirova Street, Gomel, 246653, Republic of Belarus
E-mail: yurafromgomel@gmail.com

Li Jinwei
Daihatsu Motor Co., Ltd
E-mail: vjo16252@gmail.com

***Abstract*—Urban planners have been long interested in understanding how urban structure and activities are mutually influenced. What are the forces that lead to the current patens of road network and what is the influence of the latter's structure on urban activities? Network centrality measures have been traditionally utilized to uncover the structural properties of urban street networks. In this paper we are interested in examining the correlation between the centrality of street network and the intensity of human movement in urban areas. We focus on two cities and we utilize a dataset of geo-tagged tweets that can form a proxy to urban mobility and the corresponding street networks as obtained from OpenStreeMap. Our results indicate that different centrality metrics have different levels of correlations with the intensity of human movement. Furthermore, the strength of the correlation varies in the two cities examined.**

Keywords-Human Mobility; Urban Street Network; Network Centrality; OpenStreetMap

I. INTRODUCTION

Urban spaces are typically highly localized but they are globally connected [1]. In particular, the urban space consists of local patchwork, which serve some specific functionality. Nevertheless, these patchwork are linked by the urban street network into a whole at a global scale. While the structure of urban space is greatly influenced by the history of each city [2], researchers have long been analyzing its properties in order to facilitate planning functionalities, such as resource allocation and transportation planning. Human activities in urban environments, such as business and travel, are often shaped and constrained by the geographical distance to and accessibility of the resources.

The urban street network functioning as the backbone of urban space. plays a vital role in connecting urban neighborhoods together and supporting the local/global movement in/between urban areas. Its structural properties, such as centrality and accessibility, can reveal many implications on human activities. Centrality [3], which is a network-based metric measuring the structural. importance of nodes in complex networks is often utilized to capture the importance of different parts of road networks. Former studies [4, 5] indicate that the structural properties of urban road networks as captured by the betweenness centrality can explain the observed traffic flow. Another form of centrality,

DOI: 10.21307/ijanmc-2021-039

closeness centrality is shown to be highly correlated with the intensity of economic activities [6] and land use [7]. Furthermore, the aggregated human travel flow on streets is shown through simulations to be mainly shaped by the underlying street structure [8].

In this work, we conduct a study on the correlation between the centrality of the urban street network and the intensity of human movement over it using data from Pittsburgh and NYC. Our results imply that different centrality metrics correlate with the intensity of human movement at different levels. The correlation strength further differs in the two cities examined.

Roadmap: Section 2 describes our analysis set up and the dataset. Section 3 presents and analyzes our experimental results, while Section 4 concludes by also discussing the future work.

II. Experimental Setup

In this section we will introduce the network structures that capture the intensity of human movements and the urban road network as well as the data that drive their realizations in Pittsburgh and NYC.

A. Human Transition Network

In the human transition network $G_T = (U, E)$, the set of nodes U is a collection of non-overlapping areas/neighborhoods in the city under examination. Further, a directed edge e_{ij} between two areas $u1, u2 \in U$ exists if there has been observed a transition by a city-dweller from *u1* to *u2.* The definition of *ui*s can be arbitrary (e.g., municipal neighborhood borders, grid etc.). In our analysis, we divide the whole city (10^2 miles rectangle area considered around the center of each city) into 400 neighborhood areas, each one of 0.5 miles^2

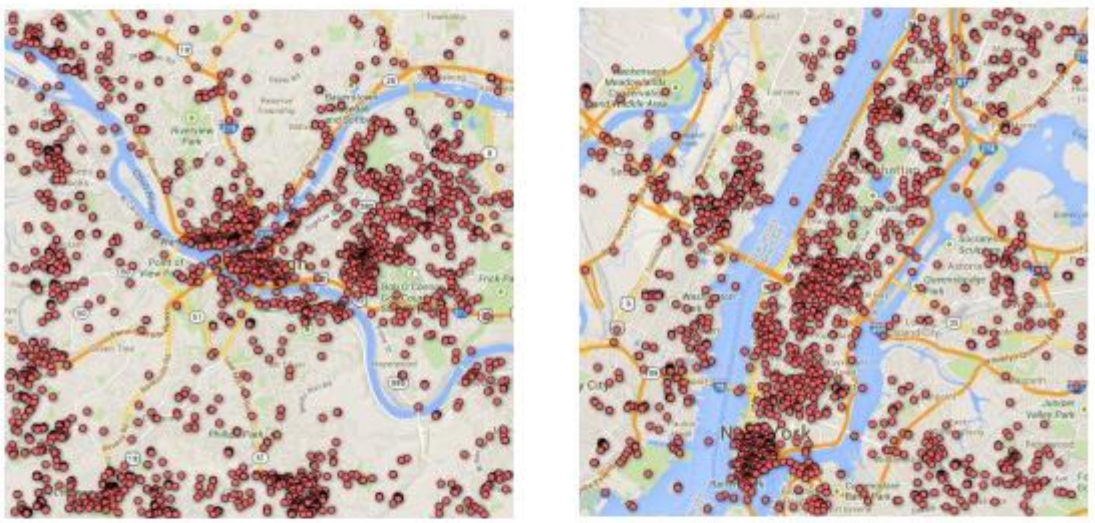
In order to obtain the structure of G_T for both cites we use geo-tagged social-media user-generated content. In particular, we collect Tweets using Twitter's streaming API from Jul 15 to Nov 15 2013. Each tweet has a tuple format<*user Id, place Id, time, latitude, longitude*>. In total, we have 526,799 geo-tagged tweets in Pittsburgh and 3,715,016 in NYC. Figure 1 presents the distribution of tweets in two cities examined. Using these data, we generate edge (transition) *eij* $\in E$ if the same Twitter user has generated two consecutive tweets in locations *li* $\in$ *ui* and *lj* $\in$ *uj* within a predefined time interval Δt and the distance between these two locations is greater than a threshold Δd. In our experiment, we set Δt = 4 hours and Δd = 10m. Finally, we have 172,887 such pairs in Pittsburgh and 961,671 in NYC. Note that the above definition allows for self-edges in *GT*. We can also annotate every edge *eij* with a weight, which captures the number of transitions between the two urban areas *i* and *j*.

Centrality in GT: To capture the centrality of human movement in different neighborhoods, we calculate the PageRank [9] for each node in *GT*. In particular, we calculate a weighted PageRank score *Pi* of area *i* as:

$$P_i = \alpha \sum_j A_{ij} \frac{P_j}{k_j^{out}} + \beta_i, \qquad (1)$$

Where a=0.85 and k_j^{out} is the weighted out-degree of node *j* which counts self and outgoing edges β_i is a personalized (external) priority importance for area *i*, which is defined as the fraction of tweets taking place in area *i.*

This work will also use a second simple centrality metric for G_T, which is the number of tweets $n_{t,i}$ generated in area *i.* The latter does not incorporate mobility information, but rather captures the intensity of activity in each area.



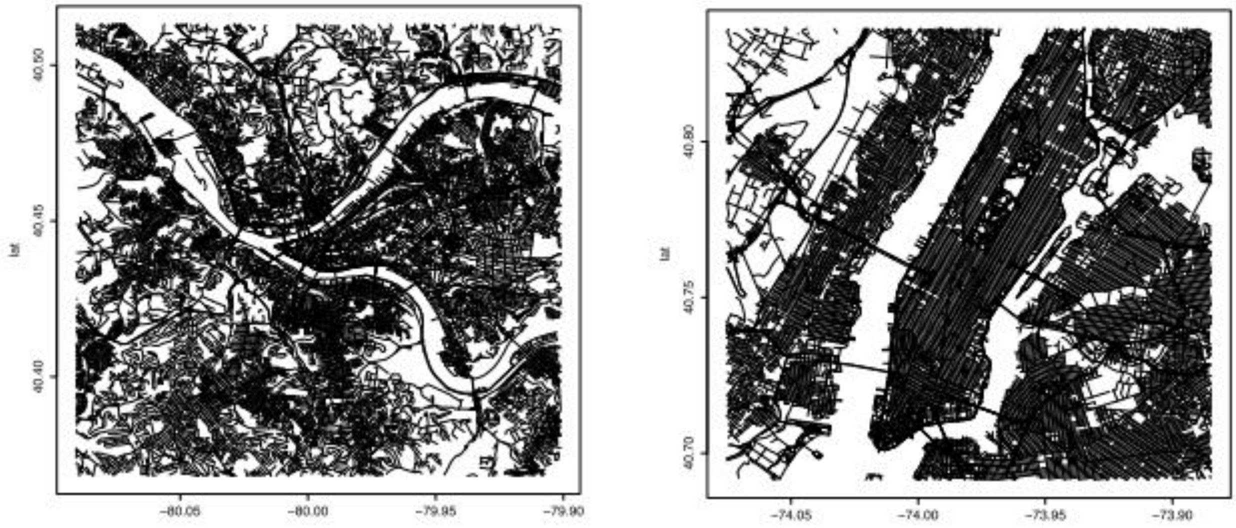
(a) Pittsburgh

(b) New York City

Figure 1. Street network in selected urban areas of two cities

B. Street Network

This paper will model the street network through a graph *Gs =(V,S)*, where the set of nodes represents the intersections in the street spatial structure and an edge $s_{ij} \in S$ represents the street segment that connects intersections *i* and *j*. We fetch the street networks from OpenStreetMap and process them using osm4arouting [1] into the G_s network format. osm4routing provides additional metadata such as the coordinates of each intersection, the length of each street segment and accessibility flags for each street segment in two directions(e.g accessibility by car, foot, bike etc.). Figure 2 further gives a visualization of the street networks in both cites.



(a) Pittsburgh

(b) New York City

Figure 2. Street network in selected urban areas of two cities

Centrality of Street Network: For a road network with *n* nodes and *m* edges, we calculate three well-established measures of node centrality: closeness centrality betweenness centrality C^b and straightness centrality C^s, C_i^c caps the accessibly of node and is defined as [3]:

$$C_i^c = \frac{n-1}{\sum_{j=1,j\neq i}^{n} d_{ij}} \tag{2}$$

Where d_{ij} is the shortest path length between nodes *i* and *j*. C_i^b quantifes to what extent node I serves as a "broker" betteen nodes, is formally defined as [3]:

$$C_i^b = \frac{1}{(n-1)(n-2)} \sum_{s=1;t=1;s\neq t\neq i}^{n} \frac{n_{st}^i}{n_{st}} \tag{3}$$

Where, n_{st} is the number of shortest paths been nodes *s* and *t* while n_{st}^i is the number of such shortest paths that traverse node *i*. C_s measures the extent to which node *i* can be reached directly, on a straight line, from all other nodes, which is defined as [6]:

$$C_i^s = \frac{1}{n-1} \sum_{j=1;j\neq i}^{n} \frac{d_{ij}^{Eucl}}{d_{ij}} \tag{4}$$

Where, d^{Eucl} is the Euclidean distance between nodes *i* and *j*.

Finlly, we calculate three global and nine local indices of street centralities. The global in indices, $C^c glob$, $C^b glob$ and $C^s glob$, are calculated using the whole road network. We also consider the local version of centralities $C^c local,d$, $C^b local,d$ and $C^s local,d$, where we compute the centrality of node *i* considering only the nodes that are within a radius *d*.

C. Analysis setup

[1] https://github.com/Tristramg/osm4routing

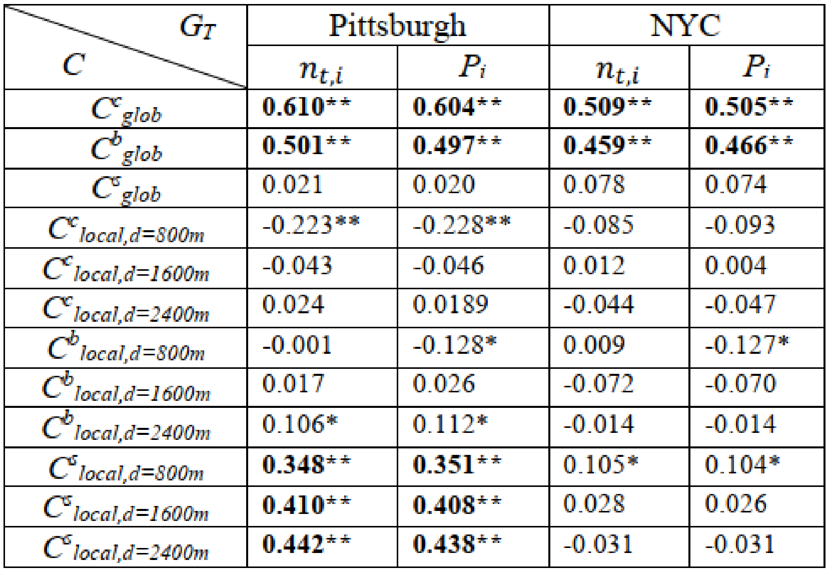
Our goal is to examine the relation between the central areas in a city as captured through the mobility of people, and the central areas of the city as captured through the street network. For that, we will utilize the Spearman's rank correlation coefficient ρ. In particular, the first variable for this correlation will be the PageRank centrality *Pi* of nodes $i \in U$ (as well as *nt,i*). However, the centrality values that we got from the street networks are defined on a different set of nodes(set *V*). Thus, we will use a spatial mapping $\Phi : V \rightarrow U$ utilizing the lat/lon coordinates we have for every $v \in V$. With Φ in place, the second variable for calculating ρ will be the average road network centrality. $\overline{C}_v^*$ of all nodes $v \in V$ that map to $i \in U$, that is, $\Phi(v)=i$.

III. Results and Analysis

We take the urban street network as an directed network without consideration of the traffic accessibility in two directions. Table 1 presents the correlation results for Pittsburgh and NYC. We can see that the global closeness centrality C^c_{glob} and betweenness centrality C^b_{glob} highly correlate with the intensity of human movement in both environments. This suggests that center areas in urban cities tend naturally to be more accessible from/to other places (higher C^c_{glob}) and thus function as city "hubs"(higher C^b_{glob}). In contrast the global straightness centrality $C^s glob$, local closeness centrality $C^c local,d$ and local betweenness centrality $C^b local,d$ present no significant positive correlations. However, the straightness centrality $C^s local,d$ shows an interesting urban difference with a significant level of correlation in Pittsburgh but not in NYC. This is more likely due to the difference of urban space structures or travel patterns between the two cities.

Further analysis is needed to sort out the exact source of this difference

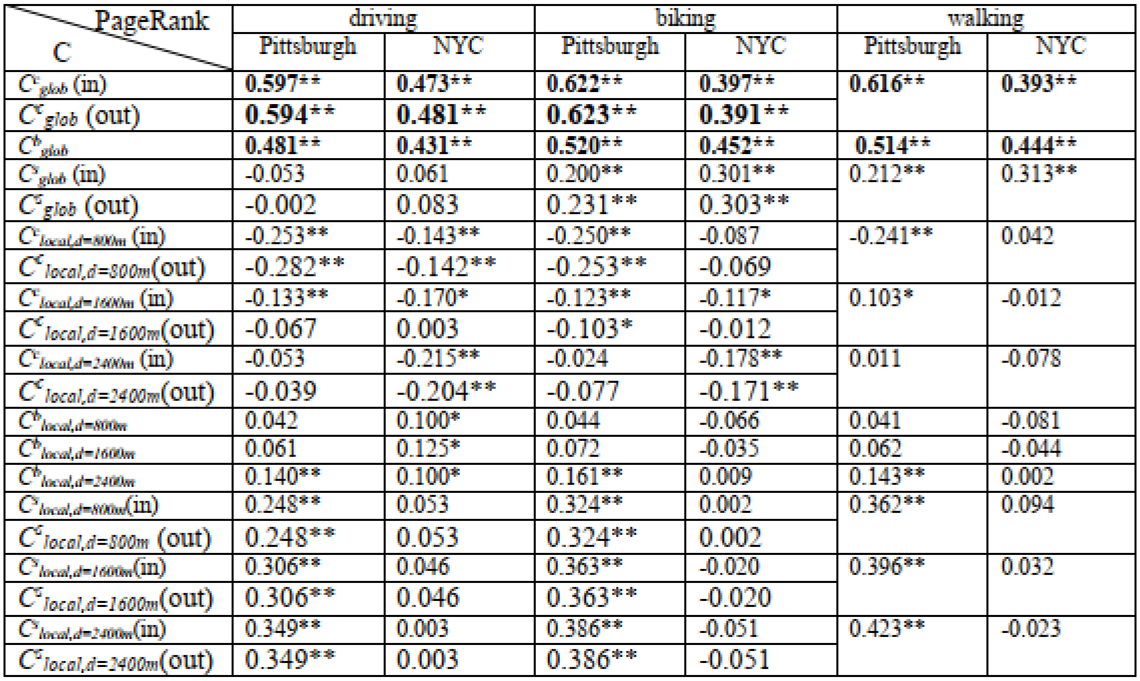
TABLE I. CORRELATION P(*INDICATES A P-VALUE<0.05; ** INDICATES P-VALUE<0.01)BETWEEN THE STREET CENTRALITY AND THE INTENSITY OF HUMAN MOVEMENT.



G_T / C	Pittsburgh		NYC	
	$n_{t,i}$	P_i	$n_{t,i}$	P_i
C^{c}_{glob}	**0.610****	**0.604****	**0.509****	**0.505****
C^{b}_{glob}	**0.501****	**0.497****	**0.459****	**0.466****
C^{s}_{glob}	0.021	0.020	0.078	0.074
$C^{c}_{local,d=800m}$	-0.223**	-0.228**	-0.085	-0.093
$C^{c}_{local,d=1600m}$	-0.043	-0.046	0.012	0.004
$C^{c}_{local,d=2400m}$	0.024	0.0189	-0.044	-0.047
$C^{b}_{local,d=800m}$	-0.001	-0.128*	0.009	-0.127*
$C^{b}_{local,d=1600m}$	0.017	0.026	-0.072	-0.070
$C^{b}_{local,d=2400m}$	0.106*	0.112*	-0.014	-0.014
$C^{s}_{local,d=800m}$	**0.348****	**0.351****	0.105*	0.104*
$C^{s}_{local,d=1600m}$	**0.410****	**0.408****	0.028	0.026
$C^{s}_{local,d=2400m}$	**0.442****	**0.438****	-0.031	-0.031

This research further consider the urban street network as a directed graph based on the direction accessibility for three types of movements including driving biking and walking. In this case, there are two different calculation or closeness and straightness centrality based on two types of shortest paths between nodes. The first one is outgoing shortest path d_{ij}^{out}, with the direction starting from node *i* to node *j*. The second is incoming shortest path d_{ij}^{in} with direction into node from node *j*, capturing how easily a traveler can access node *i* from other locations in the city. Therefore, we have *in* and *out* closeness and straightness centrality based on these two types of shortest path calculations. Table 2 presents the correlation between the centrality of directed street network and the PageRank score of neighborhood areas (results for $n_{i,t}$ are omitted due to space limitations). Compared to Table 1, we do not observe significant differences when considering the directed networks. This might be due to the fact that the transition network G_T essentially captures the starting and ending point of a movement, ignoring the actual path followed and/or due to the high similarity of the different directed network structures. Nevertheless, there is still some significant change for global straightness centrality when considering directed street network-especially for biking and walking-which might be attributed to the fact that for these "slow modes" of transportation short geometric distance is

TABLE II. CORRELATION RESULTS BY CONSIDERING THE ROAD NETWORK AS A DIRECTED NETWORK BASED ON THE ACCESSIBILITY OF DRIVING, BIKING AND WALKING IN EITHER DIRECTIONS.



PageRank / C	driving		biking		walking	
	Pittsburgh	NYC	Pittsburgh	NYC	Pittsburgh	NYC
C^c_{glob} (in)	**0.597****	**0.473****	**0.622****	**0.397****	**0.616****	**0.393****
C^c_{glob} (out)	**0.594****	**0.481****	**0.623****	**0.391****		
C^b_{glob}	**0.481****	**0.431****	**0.520****	**0.452****	**0.514****	**0.444****
C^s_{glob} (in)	-0.053	0.061	0.200**	0.301**	0.212**	0.313**
C^s_{glob} (out)	-0.002	0.083	0.231**	0.303**		
$C^c_{local,d=800m}$ (in)	-0.253**	-0.143**	-0.250**	-0.087	-0.241**	0.042
$C^c_{local,d=800m}$(out)	-0.282**	-0.142**	-0.253**	-0.069		
$C^c_{local,d=1600m}$ (in)	-0.133**	-0.170*	-0.123**	-0.117*	0.103*	-0.012
$C^c_{local,d=1600m}$(out)	-0.067	0.003	-0.103*	-0.012		
$C^c_{local,d=2400m}$ (in)	-0.053	-0.215**	-0.024	-0.178**	0.011	-0.078
$C^c_{local,d=2400m}$(out)	-0.039	-0.204**	-0.077	-0.171**		
$C^b_{local,d=800m}$	0.042	0.100*	0.044	-0.066	0.041	-0.081
$C^b_{local,d=1600m}$	0.061	0.125*	0.072	-0.035	0.062	-0.044
$C^b_{local,d=2400m}$	0.140**	0.100*	0.161**	0.009	0.143**	0.002
$C^s_{local,d=800m}$(in)	0.248**	0.053	0.324**	0.002	0.362**	0.094
$C^s_{local,d=800m}$ (out)	0.248**	0.053	0.324**	0.002		
$C^s_{local,d=1600m}$(in)	0.306**	0.046	0.363**	-0.020	0.396**	0.032
$C^s_{local,d=1600m}$(out)	0.306**	0.046	0.363**	-0.020		
$C^s_{local,d=2400m}$(in)	0.349**	0.003	0.386**	-0.051	0.423**	-0.023
$C^s_{local,d=2400m}$(out)	0.349**	0.003	0.386**	-0.051		

IV. DISCUSSION AND FUTURE WORK

In this paper we examined the correlations between the centrality of street networks with the intensity of human movement in urban areas and we found that the correlation level differs with different centrality metrics, of which some further depend on different cities. Our work provides an illuminating way to study the relationship between urban structure and human movement in a large-scale way.

We would like to emphasize that our analysis methods may suffer from a variety of biases. For example, we examine the correlation by aggregating the road network centrality and human movement in each neighborhood area, while a microscopic study might give a different view. Also, the rectangle urban area we pick may introduce edge effects on the correlation results. Furthermore, the large-scale available dataset used here may have some noises and biases. For instance, the street networks in OpenStreetMap might not that accurate especially for cities that are not that popular, since all the information is crowdsourced y the public. Also, the nature of voluntarily sharing may only give a partial information of human movement captured by geo-tagged tweets, of which the quality depends on many other factors, such as demographic biases, spam tweets and fake location information.

In the future, we plan to examine the levels of correlation by considering the temporal and contextual information of human movement such as the time and type. Furthermore, we aim to examine the centralities of a directed road network by considering the accessibility of different transportation modes (e.g., driving, biking and walking) in two directions on the street. For network centralities, we want to further investigate other practical factors, such as the max flow on a street (number of available lanes), the fastest path and the density/type of resources surrounding a street intersection.

REFERENCES

[1] Bill Hiller. Alasdair Tumer, Tao Yang, and H-T Park Metric and topo-geometric properties of urban street networks: some convergences, divergences and new results. Journal of Space Syntax Studies, 2009.

[2] J.P. Rodrigue. Transportation and the urban form. Chapter 6, The Geography of Transport Systems, 3rd Ed., 2013.

[3] Mark Newman. Networks: an introduction. Oxford University Press, 2009.

[4] Aisan Kazerani and Stephan Winter. Can betweenness centrality explain traffic flow. In Proceedings of the 12th AGILE International Conference on GIS, 2009.

[5] lan XY Leung, Shu-Yan Chan, Pan Hui, and Pietro Lio. Intra-city urban network and traffic flow analysis from gps mobility trace. arXiv preprint arXiv. 1105.5939, 2011.

[6] Sergio Porta, Vito Latora, Fahui Wang, Salvador Rueda, Emanuele Strano, Salvatore Scellato, Alessio Cardillo, Eugenio Belli, Francisco C'ardenas, Berta Cormmenzana, et al. Street centrality and the location of economic activities in Barcelona. Urban Studies, 49(7):1471-1488, 2012.

[7] Francesco Battaglia, Giuseppe Borruso, and Andrea Porceddu. Real estates ban centra, econome act. a s analysis on e ity of snon() Computational Science and Its Applicanons-ICCSA 2010, pages 1-16. Springer, 2010.

[8] Bin Jing and Tao Jia. Agent-based simulation of human movement shaped by the underlying street structure. International Journal of Geographical Information Science, 25(1):51-64, 2011.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pager-ank citation ranking Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

Development of Blind Deblurring Based on Deep Learning

Shi Kecun
School of Computer Science and Engineering Xi’an Technological University
Xi’an, China
E-mail: 528548445@qq.com

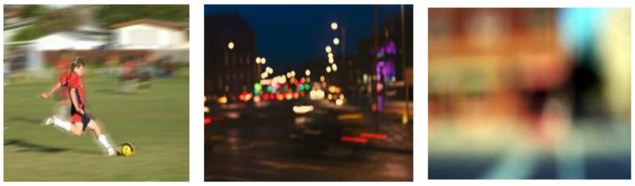
Zhao Li
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: 332099732@qq.com

***Abstract*—The rapid development of computer vision has greatly promoted the development of camera in the fields of target detection, remote sensing analysis, target recognition and so on. As the carrier of information exchange in contemporary society, image contains a large number of information elements. When taking an image, it will blur the image and destroy many photos due to various factors, resulting in the inability to obtain relevant information fr- om the image. Restoring a clear image from the blurred image is a hot spot in the field of computer vision and image processing. This paper expounds the causes of fuzziness in detail, summarizes and combs the current blind deblurring methods and research status based on deep learning, makes a detailed overview of the development of deblurring network based on deep learning from three aspects: convolution neural network, cyclic neural network and generation countermeasure network, and summarizes the advantages and disadvantages of various methods, Some networks are compared, and the problems of feature extraction, evaluation index and data set construction in deblurring are analyzed. Finally, the development trend of image motion blur restoration technology is prospected.**

Keywords-Deep Learning; Neural Networks; Computer Vision; Deep Learning; Blind Deblurring Neural Network

I. INTRODUCTION

With the advent of the intelligent era, the ways to obtain images are more extensive and convenient. Images have also become an important means for people to transmit information every day. Image restoration is the task of restoring clean images from degraded versions. Typical examples of degradation include noise, blur, rain, fog, etc. This is a highly ill posed problem because there are infinite feasible solutions. However, with the wide range of image acquisition methods, the image quality also decreases. There are many reasons for image blur. According to the formation conditions, the image blur caused by photography can be divided into motion blur, defocus blur and Gaussian blur, as shown in the figure. Among them, motion blur is the main reason for image degradation, Motion blur is also the most common kind of blur in obtaining pictures in life and one of the research hotspots. At the moment of obtaining images, the image quality degradation caused by the relative movement between the camera and the target object is called motion blur. In addition to motion blur, there are Gaussian blur and defocus blur, as shown in Figure 1. With the development of society, the rapid growth of consumer digital photography makes the camera jitter in motion blur extremely prominent. Especially with the popularity of small high-resolution cameras, these cameras are light and difficult to maintain sufficient stability. If the camera jitter occurs in the image for any reason, this moment will be “lost”. This is also an important problem that has plagued photography lovers for a long time. How to avoid image degradation and improve image quality has always been an urgent problem in the field of image processing .



a. Motion blur b. Gaussian blur c. Defocus blur

Figure 1. Different fuzzy types

Motion blurring of image plays an important role in the field of daily public safety. It is used in the field of highway safety for capturing the electronic eyes of illegal vehicles and monitoring the suspect. From a mathematical point of view, motion blur can be regarded as the result of convolution between clear image and fuzzy kernel. In actual scenes, there is usually random noise, and its mathematical model can be expressed as:

$$B = I \otimes K + N \tag{1}$$

Where B represents blurred image, I represents clear image, K represents point spread function, represents convolution, and N represents additive noise.

II. Fuzzy Reduction Method

The restoration of blurred image belongs to image restoration. The traditional deblurring methods are divided into blind deblurring and non blind deblurring according to whether the fuzzy kernel is known. For different fuzzy types, the forms of fuzzy kernel are also different. If the fuzzy kernel is known, the process of restoring a clear image is called non blind deblurring. Non blind deblurring is directly calculated from the known fuzzy kernel and fuzzy image. If the fuzzy kernel is unknown, it is called blind deblurring. Blind deblurring needs to estimate the fuzzy kernel and clear image at the same time. Only after the accurate fuzzy kernel is estimated can the clear image be restored. If the fuzzy kernel is not estimated accurately, it will directly affect the quality of the restored image. According to the prior knowledge of the blurred image, a clear image close to the real image is reconstructed from one or more blurred images. Literature [1, 2] proposed a method based on probability statistics to estimate the fuzzy kernel. The fuzzy kernel is usually inconsistent. Different pixels in a frame usually correspond to different fuzzy kernels, so it is a serious ill conditioned problem to find the fuzzy kernel corresponding to each pixel. This kind of method only has certain deblurring effect on specific images. The mathematical model is complex, the calculation efficiency is low, it is greatly affected by noise and has high requirements for fuzzy kernel estimation; Or the robustness of the algorithm is not very strong, so it can not adapt to some different data sets, and it is difficult to adapt to the fuzziness caused by some different factors. Simplified assumptions on fuzzy models usually hinder their performance in real word examples. In real word examples, fuzzy is much more complex than modeling and entangled with the image processing pipeline in the camera.

In addition, recent machine learning based methods also rely on synthetic fuzzy data sets generated under these assumptions. This makes the traditional deblurring method unable to remove the fuzzy kernel, which is difficult to approximate or parameterize (such as object motion boundary). Some learning based methods are also proposed for deblurring blurred images. Recent work has begun to use end-to-end trainable networks for image [3] and video [4, 5] deblurring. Non uniform blind deblurring of general dynamic scenes is a challenging computer vision problem, because blur comes not only from the motion of multiple objects, but also from camera jitter and scene depth change.

A. Method of Convolutional Neural Network Based on Deep Learning

In order to limit the solution space to effective natural images, the existing restoration technologies [6, 7, 8] explicitly use the image priors made by hand through empirical observation. However, designing such a priori knowledge is a challenging task and often can not be popularized. In order to improve this problem, deep learning methods begin to be more applied to image processing. Recently, the most advanced method [17, 44,] adopts convolution neural network (CNN). After 2016, the early CNN based image deblurring method usually uses CNN as a fuzzy kernel estimator to construct a two-stage image deblurring framework, such as CNN based fuzzy kernel estimation stage and kernel based deconvolution stage. Chakrabarti uses CNN to estimate the fuzzy kernel [12], obtains the fuzzy kernel, and then uses the non blind deblurring algorithm to deblurring. Schuler et al. [13] trained a depth network to estimate the fuzzy kernel, and then used the traditional non blind deconvolution method to restore the potentially clear image. Sun et al. [14] used convolutional neural network to estimate the image fuzzy kernel, and then used the estimated fuzzy check image for deblurring restoration. Although this method uses the blind deblurring algorithm to restore the image, it still uses the idea of non blind deblurring after estimating the fuzzy kernel to deconvolute the image.

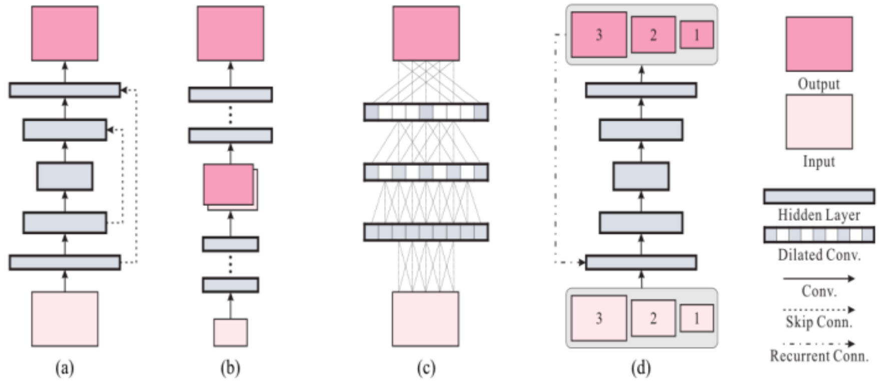


Figure 2. Different CNN for image processing. (a) U-net or Codec network. (b) Multiscale or cascade refinement network. (c) Extended convolution network. (d) Scale recursive network (SRN).

This leads to the slow operation of the algorithm and the restoration result depends on the fuzzy kernel estimation. However, this method applies the convolution neural network to image deblurring, which lays the foundation for subsequent methods based on this. Li et al. [15] proposed a new convolution structure called "hole convolution", and its kernel is calculated by a rectangular rectangular ring, The experimental results show that this method can effectively restore the image;Liu et al. [16] proposed a two-stage deblurring module to restore the blurred image of the dynamic scene based on the high-frequency image. Firstly, the residual image is thinned by the coding network, and then the thinned residual image is combined with the input blurred image to obtain the latent image, and further proposed a coarse to fine framework based on the fuzzy processing module. Noroozi uses multi-scale CNN for end-to-end training [17], which does not need to estimate the fuzzy kernel and belongs to blind fuzzy. The basic idea of the blind deblurring method based on CNN is to take the blurred image and the corresponding clear image as training samples and input them into the convolution neural network for training. After the training, the optimized network model is obtained. When in use, the blurred image is taken as the network input, and the network output is the deblurring image. Different image processing is based on different convolution networks, as shown in Fig. 2.

On the other hand, the recent image deblurring method based on CNN aims to directly understand the complex relationship between blurred and clear image pairs in an end-to-end way. NAH et al. [18] proposed using multi-scale convolutional neural network to defuzzify and created the most widely used GoPro data set at present. The "end-to-end" training method is adopted, which has good model effect and operation speed. It can directly recover the latent image without assuming any limited fuzzy kernel model. In particular, the multi-scale structure is designed to imitate the traditional coarse to fine optimization method. Unlike other methods, this method does not estimate explicit potential errors. Therefore, artifacts due to kernel estimation errors will not be generated. Secondly, the proposed model is trained by multi-scale loss. As shown in the figure, the model is suitable for the structure from coarse to fine, which greatly enhances the convergence. The multi-scale system uses the improved residual network structure, as shown in Figure 3, to achieve a deeper architecture. In addition, the results are further improved by using counter loss [19]. Because the loss term optimizes the result and makes it similar to the ground truth, it even restores the extremely complex occlusion area of the fuzzy kernel, and has made significant improvements in the deblurring of dynamic scenes. A large number of experimental results show that the performance of this method in qualitative and quantitative evaluation is much better than the latest dynamic scene deblurring method.

B. Method Based on Cyclic Neural Network

The cyclic neural network is used in the image motion fuzzy restoration method. This method innovatively uses the cyclic neural network processing data with sequence and time dependence for image processing.

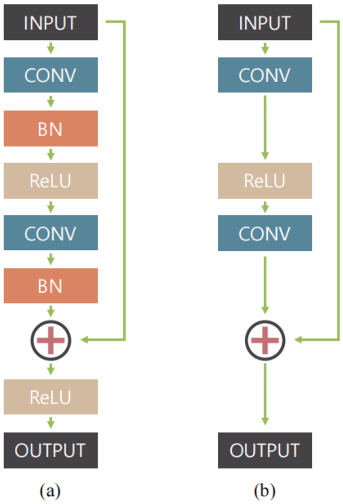


Figure 3. (a) The original remaining network building blocks. (b) Building blocks of the modified network by NAH et al.

Tao et al. [20] proposed a scale cyclic neural network (SRN deblurnet) according to the strategy of gradually recovering clear images with different resolutions in the pyramid, Compared with other methods, SRN has simpler structure, fewer parameters and easier training than other networks. Resblock network is also inspired by the recent success of encoder decoder structure used in various computer vision tasks, and explores an effective method to adapt it to image deblurring tasks. In SRN network, the direct application of the existing encoder decoder structure can not produce the best results. The encoder decoder resblock network of Tao et al. Amplifies the advantages of various CNN structures, produces the feasibility of training and produces a very large receptive field, which is very important for large motion deblurring. Experiments show that the end-to-end depth image deblurring framework can greatly improve the training efficiency by using the cyclic structure and combining the above advantages.

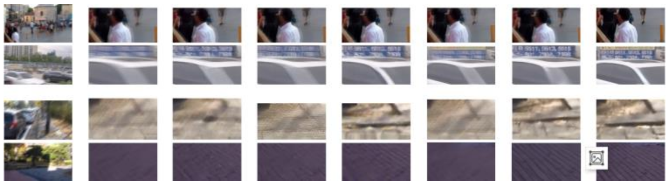
Zhang et al. [21] proposed a depth hierarchical multi patch network based on spatial pyramid matching, which processes fuzzy images through a fine to rough hierarchical representation, It runs 40 times faster than the previous multi-scale method. Zhang et al. [22] proposed a spatial variation neural network composed of three deep convolution neural networks (CNNs) and a cyclic neural network (RNN). RNN is used as a deconvolution operator to deconvolute the feature map extracted from the input image by a neural network. This method has good performance, speed and model size.

C. Defuzzification Method Based On Generation CounterMeasure Network

Most of the traditional deblurring methods based on convolutional neural network have a series of problems, such as the color of the output image is unnatural, the texture features are not rich enough, the image is too smooth and so on. Confrontation network (GAN) is also gradually applied to the field of image deblurring because it can retain texture details and generate realistic images. In 2014, goodflow et al. Proposed a groundbreaking generation confrontation network (GAN) that can show strong ability in computer vision tasks [23].

The images processed by the generation confrontation network are very close to clear images, It can not even be distinguished with the naked eye [24-25].However, the generator should be constrained when using the generated countermeasure network, because once the network is too free, it will lead to instability, and it is difficult to learn the mapping relationship between the input image and the target image. If L1 or L2 constraints are directly established between the output of the generator and the corresponding target image at the pixel level, the generated image will become too smooth and vulnerable to image noise. With the increasingly prominent application of generative countermeasure network in the image field, ledig et al. [26] proposed a generative countermeasure network (GAN) for image super-resolution (SR). Its deep residual network can restore realistic texture, and the results are obtained through the mean score (MOS) test. Using srgan can improve the quality of perception, this method can be well applied to the field of image deblurring.

In 2018, kupyn et al. [27] first removed the blur of camera jitter according to the conditional counter- measure network, and then proposed a kernel free blind motion defuzzification learning method to make up for the previous shortcomings. The conditional countermeasure network deblurgan optimized by using multi-component loss function, but this method did not consider the impact of different feature layers in the perception network on the perception loss, The detail of the restored image is still smooth. Meanwhile, in 2019, kupyn et al. [28] introduced the feature pyramid network into deblurgan for the first time and proposed a new end-to-end generation countermeasure network deblurgan-v2 for single image motion deblurgan, which greatly improved the efficiency, quality and flexibility of deblurgan. Table 2 shows more depth learning methods for blind image deblurring. Then the network is reproduced, and the deblurring image is shown in Figure 4.



(a)Blurred Image (b)Blurred patch (c) MS-CNN (d)DeblurGAN (e)SRN (f)DelurGAN-V2 (g)MPRNet (h)MIMO-UNet

Figure 4. Visual comparison of image deblurring results of GoPro test set [13]. Patches blurred by key points are displayed in (b), while patches magnified from deblurring results are displayed in (c) - (h).

III. KEY PROBLEMS OF IMAGE MOTION BLUR RESTORATION

A. Feature Extraction

Image recognition is actually a classification process. In order to identify the category of an image, we need to distinguish it from other different categories of images. This requires that the selected features not only can well describe the image, but also can well distinguish different types of images. We want to select the image features with small difference between similar images (small intra class spacing) and large difference between images of different categories (large class spacing), which we call the most discriminative feature. In addition, prior knowledge plays an important role in feature extraction. How to rely on prior knowledge to help us select features is also a problem that will continue to be concerned later. The research process and ideas of traditional feature extraction methods are very useful，Because these methods have strong interpretability, they provide inspiration and analogy for designing machine learning methods to solve such problems. The existing convolutional neural network is similar to these feature extraction methods, because each filter weight is actually a linear recognition pattern, which is similar to the boundary and gradient detection of these feature extraction processes. At the same time, the role of pooling is to coordinate the information of a region, which is similar to the feature integration (such as histogram) after these features are extracted. Through experiments, it is found that the first few layers of convolution network are actually doing edge and gradient detection. However, in fact, when the convolution network was invented, there were no such feature extraction methods.

B. Evaluation Method

The evaluation of image quality can be divided into subjective evaluation and objective evaluation. Subjective evaluation mainly evaluates human visual senses, while objective evaluation uses an evaluation standard to compare image quality. The commonly used objective evaluation methods are peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [29]. The PSNR reflects the distortion degree of the estimated image and the original clear image. Generally speaking, the larger the peak signal-to-noise ratio, the better the image restoration effect. Its expression is:

$$\mathrm{PSNR} = 10\log 10\left[\frac{\mathrm{M}\times\mathrm{N}\times 255^2}{\|\mathrm{X}-\hat{\mathrm{X}}\|_2^2}\right] \tag{2}$$

M and N represent the size of the row and column of the image respectively, and X represents the original clear image, $\hat{x}$ represents the estimated image. Due to research findings, PSNR is sometimes inconsistent with human visual evaluation. Therefore, SSIM is adopted to further improve the evaluation criteria. Its expression is:

$$\mathrm{SSIM} = \left[l(x,\ y)\right]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{3}$$

TABLE I. COMPARISON OF CHARACTERISTICS OF MAINSTREAM DATA SETS

Data Set	Construction Method	Advantages and Disadvantages
Levin etc.	Algorithm simulation fuzzy kernel	Easy to obtain; It is easy to obtain without considering local fuzziness;
Kupy etc.	Simulated trajectory	Easy to obtain; Only the motion in two-dimensional space is simulated, and the real three-dimensional space is not considered
Kohler etc.	The motion track is captured by 6D camera	The motion trajectories in three-dimensional space are collected; Lens distortion, depth of field variation, etc. are not considered
GOPRO etc.	Take the average value for continuous shooting by high-speed camera	Closer to the real fuzzy situation; The acquisition process is troublesome and the data scene is single
Lai etc.	Real acquisition	Completely real fuzzy pictures; There is no corresponding clear image, which is often used as a test set

$$l(x,y) = \frac{2u_x u_y + c_1}{u_x^2 + u_y^2 + c_1} \tag{4}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{5}$$

$$s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{6}$$

SSIM is a comprehensive map image evaluation index, which evaluates the image from brightness, contrast and structural similarity respectively, where u_x and u_y represent the mean value of image X and Y respectively, X and Y represent the difference between image X and Y respectively, and XY represents the co difference between image x and y. SSIM measures the similarity of two images, and its value is between 0 and 1. The closer it is to 1, the higher the similarity is the better the restoration result is.

C. Data Set Construction

*Most of the blurred images in the traditional data set are blurred by some fixed cores, which i*s difficult to imitate the natural blurred images. When we use the algorithm in machine learning to deal with some problems, the quality of the data set will directly affect the results of our algorithm. Therefore, high-quality data sets play an important role in our follow-up research. The most direct way to obtain the data set is to directly capture the image in the real scene, the data set of Lai et al. [30]. The image obtained in this way only has fuzzy image and no corresponding clear image, so it can only be used for testing and can not be used for the data set of network training. The restoration of fuzzy images through neural networks requires pairs of fuzzy clear image pairs for network training, but such data sets are difficult to obtain in the real world. The manual use of paired data sets can not ensure the consistent content of clear images and fuzzy images, and the fuzzy images synthesized by the algorithm can not contain various complex factors in the real environment. It may perform well when used for network training, but the effect is really unsatisfactory when using real fuzzy images.

IV. Trends and Prospects

Image deblurring has attracted more and more attention in the field of image processing. It not only has important theoretical significance, but also has urgent needs in practical application. Both theoretical research and practical application have made more achievements and progress, but there are still some aspects to be improved in the future waiting for us to improve and solve.

A. Update of Data Set

In deep learning, the quality of data sets directly affects the subsequent experimental results. The quality and updating of data sets are of great significance to image deblurring. Among them, only 2103 pairs of training pictures and 1111 pairs of test pictures are used in GoPro dataset, which is the most widely used and the largest.

Compared with datasets in other fields of computer vision, especially the Imagenet dataset contains 14197122 pictures, it is very different. GoPro data sets only expand the number of data sets, the diversity of data sets obtained is not enough, and the scene is too single, and some even show that motion blur is not particularly obvious.

Similarly, some other data sets synthesized by algorithms show less data and are not sufficient in demonstration. Therefore, in the current situation, we need to enrich and update the data, not only to ensure that the amount of data is sufficient, but also to fully meet the requirements of the experiment. Different from image recognition or image segmentation, it is difficult to obtain fuzzy image data set. However, for any field, data sets are the basis of researchers' development. The lack of data sets directly affects the research progress in this field. Therefore, it is urgent to propose a large-scale and new data set.

TABLE II. BLIND DEBLURRING ALGORITHM BASED ON DEEP LEARNING

Method	**Applicable Scenario**	**Mechanism**	**Advantage**	**Limitations**
Spatial variation RNN[22]	Motion blur, dynamic scene blur	The deblurring process is formulated through the wireless impulse response model	Weights can be learned from another network and different weights can be learned for different fuzzy systems	Large regional and spatial change structures need to be involved at the same time’
SRN[20]	Motion blur	New multiscale cyclic network structure	The number of trainable parameters is reduced and the training efficiency is improved	Limited to fixed data sets and training periods
DMPHN[21]	Motion blur	End to end CNN hierarchical model similar to spatial pyramid matching	The required filter is small and can be inferred quickly	Requires large GPU memory
DPSR[32]	LR blurred image	A new SISR degradation model is designed	The deep plug and play framework can deal with any fuzzy kernel	For most real images, it does not match the degradation model
BIE-RVD[33]	Motion blur	Automatic coding structure of spatiotemporal video screen based on end-to-end differentiable structure	High accuracy and fast network running speed	The task of training is complex and difficult
DDMS[34]	Motion blur	A full convolution structure with filtering transformation and characteristic modulation is constructed	Real time filtering completely eliminates multi-scale processing and large filters	Real time filtering completely eliminates multi-scale processing and large filters
deblurGAN[27] DeblurGANV2[28]	Motion blur Motion blur	The generated countermeasure network based on perceptual loss [9] (perceptual loss) constraint is used for deblurring	The restored image is more similar to the target image in semantics and closer to people's subjective evaluation of image quality	The influence of different feature layers in the perceptual network on the perceptual loss is not considered, so that the restored image details are still smooth.
Deepdeblur[18]	dynamic scene blur	End to end multiscale convolution network	Without estimating the fuzzy kernel, multi-scale CNN can restore clear images directly and flexibly	The multi-scale stacked sub network results in large amount of parameters, large consumption of video memory and great difficulty in training
SRN-deblur[20]	Blur of dynamic scene	End to end multiscale cyclic network	Multi-scale structure and parameter sharing alleviate the problem of large amount of parameters, and the learning ability is more stable	The edge is too smooth and there are artifacts

Method	Applicable Scenario	Mechanism	Advantage	Limitations
DMPHN[21]	Motion blur	The deep-seated multi-facet network based on spatial pyramid matching processes fuzzy images through fine hierarchical representation.	It can solve the problem of performance saturation and run faster than multi-scale method	It can solve the problem of performance saturation and run faster than multi-scale method
MPRnet[35]	Deblurring, rain removing and noise removing	A multi-stage progressive image restoration	It can output accurate spatial details and context information. The network structure is simple and the effect is good	The deblurring effect under the dark light line is not good
MIMO-Net[29]	Motion blur	Single encoder multiple input single decoder multiple output	Increase the network feeling field and make the training less difficult	The spatial details are lost and the texture is not clear enough

B. Algorithm Efficiency

The important factor affecting the application of the algorithm is the running speed of the deblurring algorithm. Some scenes require high real-time performance, and the efficiency of the algorithm is the first choice. To improve the timeliness of the image motion blur restoration algorithm, this algorithm can be applied to many scenes to improve the solution based on computer vision. For example, in factory production monitoring, more and more attention is paid to the use of image processing technology. In the traditional method, the articles need to stop at the monitoring point to collect images in the production process. Using the image motion deblurring algorithm with high real-time performance can collect pictures when the articles are moving, save the steps of stopping the articles, and greatly improve the production efficiency of the article production line. Therefore, improving the efficiency of the algorithm plays an important role in daily life.

C. More Objective Evaluation of Indicators

Nowadays, people's subjective feeling is very close to the widely used structural similarity index and peak signal-to-noise ratio, especially when there is uneven fuzzy motion blur in the training image, so a reference evaluation index is good enough to evaluate the processed deblurring image and the rationality of the algorithm when processing the experimental results. With the continuous development of deblurring technology, in order to achieve a fair and fair evaluation of the deblurring image, we not only attach people's perception test effect on the image at the back of the paper, but also need to get a more recognized evaluation standard in the field of deblurring.

V. SUMMARY

This paper systematically summarizes the current research status of image motion blur restoration technology, points out the key problems of the existing research, and looks forward to the future development trend and application prospect, which lays a foundation for further research.

REFERENCES

[1] Whyte O, Sivic J, Zisserman A, et al. Non-uniform Deblurringfor Shaken Images [J]. International Journal of Computer Vision, 2012, 98(2):168-186.

[2] Zhou Luoyu, Zhang Bao, Yang Yang. Estimation of parameter of defocused blurred image using Hough transform [J]. Infrared and Laser Engineering, 2012, 41(10): 2833-2837.

[3] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. Pages 3883–3891, 2017.

[4] T. Hyun Kim, K. Mu Lee, B. Scholkopf, and M. Hirsch. Online video deblurring via dynamic temporal blending network. In ICCV, pages 4038–4047. IEEE, 2017.

[5] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring. Pages 1279–1288, 2017.

[6] Weisheng Dong, Lei Zhang, Guangming Shi, and Xiaolin Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization. TIP, 2011.1

[7] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. TPAMI, 2010.1

[8] Kwang In Kim and Y ounghee Kwon. Single-image super-resolution using sparse regression and natural image prior. TPAMI, 2010.1

[9] Wu D, Zhao H T, Zheng S B. Motion deblurring method based on DenseNets [J]. J Image Graph, 2020, 25(5): 890–899.

[10] Zhang Y L, Tian Y P, Kong Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 2472–2481.

[11] Shao W Z, Liu Y Y, Ye L Y, et al. DeblurGAN+: Revisiting blind motion deblurring using conditional adversarial networks [J]. Signal Processing, 2020, 168: 107338.

[12] Chakrabarti A. A Neural Approach to Blind Motion Deblurring[C]. European Conference on Computer Vision, Springer, Cham, 2016: 221-235.

[13] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Scholkopf. Learning to deblur. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(7):1439–1451, 2016. 2

[14] Sun J， Cao W F， Xu Z B， et al. Learning a convolutional neural network for non-uniform motion blur removal[C]// IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015:769-777.

[15] Li J，Li K，Yan B. Scale-aware deep network with hole convolution for blind motion deblurring[C]//IEEE Inter national Conference on Multimedia and Expo(ICME), Shanghai，China, 2019:658-663.

[16] Liu K，Yeh C，Chung J，et al. A motion deblur method based on multi- scale high frequency residual image learning [J]. IEEE Access，2020，8：66025-66036.

[17] Noroozi M, Chandramouli P, Favaro P. Motion Deblurring in the Wild[C]. German Conference on Pattern Recognition, Springer, Cham, 2017: 65-77.

[18] Nah S, Kim T H, Lee K M. Deep Multi-scale convolutional neural network for dynamic scenedeblurring [J]. IEEE Computer Vision and PatternRecognition, 2017, 35(1):257-265.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672–2680, 2014.

[20] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8174–8182, 2018.

[21] Zhang H，Dai Y，Li H，et al. Deep stacked hierarchical multi- patch network for image deblurring [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019:5971-5979.

[22] Zhang J.Dynamic scene deblurring using spatially variant recurrent neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition，Salt Lake City, UT, 2018:2521-2529.

[23] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets[C]. Advances in Neural Information Processing Systems, 2014:2 672-2 680.

[24] Isola P, Zhu J Y, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks[C]. IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017:5 967-5 976.

[25] Sixt L, Wild B, Landgraf T. Rendergan: Generating Realistic labeled Data [J/OL]. https://openreview.net/forum?id=BkGakb9lx, 2017-01-12/ 2018-05-02.

[26] Ledig C, Theis L, Husz ár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, July 21-26, 2017. Piscataway, NJ: IEEE, 2017: 4681-4690.

[27] Kupyn O, Budzan V, Mykhailych M, et al. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks [J/OL]. https://arxiv.org / pdf / 1711.07064.pdf, 2017-11-21 / 2018-05-02.

[28] Kupyn O, Martyniuk T, Wu J, et al. Deblurgan-v2:Deblurring (orders-of-magnitude) faster and better[C]//Pro-ceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Oct. 27-Nov. 2, 2019. Piscataway, NJ: IEEE, 2019: 8878-8887.

[29] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J].IEEE transactions on image processing, 2004, 13(4):600-612.

[30] Lai W，Huang J，Hu Z，et al. A comparative study for single image blind deblurring[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016: 1701-1709.

[31] Levin A，Weiss Y，Durand F，et al. Understanding and evaluating blind deconvolution algorithms[C]//IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009: 1964-1971.

[32] Zhang K, Zuo W, Zhang L. Deep plug-and-play super-resolution for arbitrary blur kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, June 15-20, 2019. Piscataway, NJ: IEEE, 2019: 1671-1681.

[33] Purohit K, Shah A, Rajagopalan A N. Bringing alive blurred moments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Long Beach, June 15-20, 2019. Piscataway, NJ: IEEE, 2019: 6830-6839.

[34] Purohit K, Rajagopalan A N. Region-adaptive dense network for efficient motion deblurring [C]//Proceedings of the AAAI Conference on Artificial Intelligence, New York, February 7-12, 2020. Published by AAAI Press, Palo Alto, California USA, 2020, 34(7): 11882-11889.

[35] Bai Y, Jia H, Jiang M, et al. Single-image blind deblurring using multi-scale latent structure prior [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(7): 2033-2045.