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: Dec. 28, 2020

Table of Contents

Deep Capsule Network Handwritten Digit Recognition.. 1

Yuxing Tan, Hongge Yao

Improve Usability of Tourism Websites Based on Agile Strategies... 9

Tsegaye Yabsra Asefa

Health Assessment Based on D-S Evidence Theory of Equipment... 15

Tanghui Sun, Bailin Liu

Adaptively Truncating Gradient for Image Quality Assessment.. 27

Minjuan Gao, Hongshe Dang, Xuande Zhang

Discussion on Decimal Network Based on IPV9.. 34

Hu Shun, Xu Dongmei, Gao Lin

Review of Anomaly Detection Based on Log Analysis... 40

Wu Xudong

Software TLB Management Method Based on Balanced Binary Tree... 50

Chen Hongyu, Zhang Yuke, Zhao Li, Ai Jian

Data Visualization Analysis of COVID-19 Epidemic Situation... 58

Yang Tao

Conditional GAN-based Remote Sensing Target Image Generation Method.. 66

Liu Haoyang, Yu Jun, Hu Zhiyi, Gao Shouyi

Imperative to Build Network Security System and Speed Up the Future Network Construction........................ 75

Wu Aiqun, Gao Zhangxing

Deep Capsule Network Handwritten Digit Recognition

Yuxing Tan
School of Computer Science and Engineering Xi'an Technological University
Xi'an, China
E-mail: Yuxing_Tan@foxmail.com

Hongge Yao
School of Computer Science and Engineering Xi'an Technological University
Xi'an, China
E-mail: 835092445@qq.com

***Abstract*—Aiming at the weakness of CNN that is not sensitive to the changes of relative position and angle, a method of digital handwritten recognition based on deep capsule network is researched. The capsule network represents multiple attributes of an entity through a group of capsules composed of neurons, which effectively preserves the information about the position and posture of the entity. Dynamic routing algorithm makes the information interaction between capsules more clearly, and can determine the pose of the entity more accurately. While solving the shortcomings of convolutional neural networks, it also integrates the advantages of CNN and considers the relative position of it's lack, so that the recognition effect is improved. The design implements a deep capsule network, reduces the amount of trainable parameters by changing the size of the convolution kernel, expands on the original network structure, adds a convolution after the convolution layer, and a process of dynamic routing on the main dynamic routing is added, and the number of iterations is changed for experimentation, which makes the accuracy of network recognition higher on the MNIST data set.**

Keywords-Component; Deeplearning; Nerve Capsule; Deep Capsule Network; Handwritten Digit Recognition

I. INTRODUCTION

In our daily life, handwritten numbers are very common, but in many areas of work, the part about numbers is sometimes very cumbersome, such as data collection, which is a time-consuming, large amount of work. At this time, the function of handwriting recognition technology is reflected, which brings convenience and efficiency to human.

The proposal of nerve capsule comes from a assumption of Hinton[1]: instead of using a group of coarse coding or single neurons to represent the relationship between the observer and the object similar to the object's posture information, a set of activated neurons is selected to represent it. This group of neurons is called nerve capsule. One of the advantages of capsule network is that it needs less training data than convolutional neural network, but the effect is not inferior to it.

For the traditional neural network, neurons can not represent multiple attributes at the same time, resulting in the activation of neurons can only represent a certain entity. In this way, the nature of the entity will be hidden in a large number of network parameters. When adjusting the network parameters, we can not guarantee the pure motivation. It must take into account the input of all kinds of samples, so it is inevitable to adjust the parameters in a troublesome and time-consuming manner. After the application of vector neurons, we will be able to determine the existence of all the properties wrapped in a capsule, in the adjustment of parameters, such constraints will be greatly reduced, the best parameters are easy to obtain.

The design and research of artificial neural network largely borrows from the structure of biological neural network. In the field of neurolysis, a conclusion has been drawn that there are a large number of cortical dominant structures in the cerebral cortex of most primates. There are hundreds of neurons in the cortex of most primates, and there are also hierarchical structures in it. These small units can handle different types of visual stimuli well. The researchers speculate that there is a mechanism in the brain that combines low-level visual features with some weight values to construct a colorful world in our eyes. Based on this discovery in biology, Hinton suggests that it is more appropriate to try to replace the relationship between the object and the observer with a series of active neurons instead of one. So, there is the nerve capsule mentioned earlier.

DOI: 10.21307/ijanmc-2020-031

In October 2017, Sabour, Hinton and others published the topic "Dynamic Routing Between Capsules "[10] at a top-level conference on machine learning called "NIPS" and proposed Capsule network (CapsNet). This is a deep learning method that shakes the whole field of artificial intelligence. It breaks the bottleneck of convolutional neural network (CNN) and pushes the field of artificial intelligence to a new level. This paper focuses on the recognition of MNIST data set based on capsule network. MNIST[7] is a data set composed of numbers handwritten by different people.

Although handwritten digits in BP neural network[9] and convolution neural network[2][5][6][11] have a certain good recognition effect, but the emergence of capsule network brings a new breakthrough to the recognition of data sets, and has a better recognition effect, and it’s recognition accuracy greatly exceeds the convolutional neural network.

II. Related Work

The neural capsule proposed by Hinton is to implement ontology from the perspective of philosophy. The various properties of a particular entity are represented by the activity of nerve cells in an activated capsule. These attributes include the size, location, orientation and other information of the entity. From the existence of some special attributes, we can infer the existence of instances.

In the field of machine learning, the probability of entity existence is represented by the output size of independent logistic regression unit. In the neural capsule, the norm obtained by normalizing the output high-order vector represents the existence probability of the entity, and the attributes of the entity are represented by various "posture" of the vector. This reflects the essence of ontology, that is to define the existence of entity according to its various attributes.

In the research of capsule network, the working process of capsule network is closer to the behavior of human brain because of its less training data. In the aspect of white box adversarial attacks, capsule network shows strong resistance. Under the effect of the fast gradient symbol method, the accuracy can still be maintained above 70%. The accuracy of training and testing on MNIST is better than that of convolution neural network. In some practical applications, such as in specific text classification tasks, convolution capsule network can effectively improve the accuracy of feature extraction. [12]Chinese scholars have also applied the visual reconstruction method based on capsule network structure in the field of functional magnetic resonance imaging. In the intelligent traffic sign recognition, by introducing pooling layer into the main capsule layer, the super depth convolution model improves the feature extraction part of the original network structure, and uses the moving index average method to improve the dynamic routing algorithm, which improves the recognition accuracy of the network in the field of traffic sign recognition.

The capsule network first appeared in the article "Dynamic routing between capsules" published by Hinton et al. in October 2017. Based on the capsule network proposed by Sabour et al in 2017, an improved version of the new capsule system was proposed in the article "Matrix Capsules with EM Routing"[3] published in 2018.

In this system, each encapsulated capsule uses a logical unit to represent the presence or absence of an entity. A 4×4 pose matrix is used to represent the pose information of the entity. In this paper, the iterative routing method between capsule layers based on EM algorithm is mentioned. The output of the lower layer capsules reaches the higher level capsules through routing algorithm, so that the activated capsules get a group of similar pose voting. The new system is much more resistant to Lily attacks than baseline CNN. In the paper "Stacked Capsule Autoencoders"[4] published in 2019, an unsupervised capsule automatic encoder (SCAE) is introduced. By observing the neural encoders of all components, the existence and pose information of the target can be inferred, that is to say, the object can be inferred explicitly through the relationship between the components. The accuracy on SVHN[8] and MNIST datasets is 55% and 98.5%, respectively.

III. Deep capsule network

A. Structure of deep capsule network

1) Encoder structure of deep capsule network

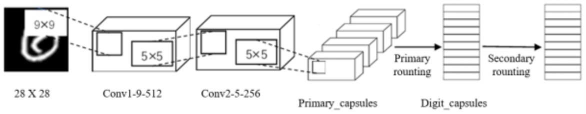


Figure 1. Network structure of deep capsule

Conv1: Standard convolution. It is dedicated to extract some low-level feature information from the input image. The preprocessing data layer of capsnet is to convert the brightness of pixels in the input layer into local feature output. The input image of this layer is 28×28, with 256 convolution kernels with step size of 1 and size of 9×9. After convolution, the output is a three-dimensional array. By reshaping the array, the appropriate feature vector of position information is constructed for each dimension. The final output is a tensor of 20×20×256.

Conv2: Standard convolution layer, including 256 convolution cores with step size of 1 and size of 5×5, input tensor of 20×20×512 and output of tensor of 16 ×16×256.

Primary capsule: primary capsule layer, also known as the primary capsule layer, is in a low-level stage, multidimensional entities are described in capsnet from the perspective of "inverse graph". It is a reverse rendering process, that is, this layer can combine the low-level features detected by the previous layer. This layer is still committed to extracting feature information, so it still belongs to convolution layer. The object of convolution is changed from single neuron to capsule with larger granularity, which is the difference between convolution network and convolution network. The primary capsule layer is the convolution layer of "capsule version". This stage is also where the capsule really begins. This layer consists of 32 main capsules, each of which contains 8 convolution kernels of 9×9 ×256 with step size of 2. According to the above, the tensor of 6× 6×8×32 is obtained by inputting 20× 20×256 tensors in this layer.

Digitcaps: Digital capsule layer, also the full connection layer of capsule network. Using a fully connected topology, the capsules in this layer will connect all outputs of the previous primary capsule layer. Because this paper finally realizes the recognition of 0-9, so there are 10 capsules in this layer. The norm of each activation vector represents the probability of each classification and is used to calculate the classification loss. The input received by this layer is the tensor of 6×6 ×8×32 of the output of the previous layer, and the output is a matrix of 16×10.

Finally, the capsule network is compared with the improved deep capsule network as shown in the following table 1:

2) Decoder structure

The decoder structure in this paper is the same as capsnet, as shown in the figure. The goal of capsnet model optimization is to calculate the edge loss for each number to allow multiple numbers to exist at the same time. In addition, capsnet can reconstruct the input image based on the instantiation parameters obtained by previous processing. In the training process of image reconstruction, only the activated capsules are allowed to participate in the adjustment of three-level fully connected network at each time. The structure mainly responsible for reconstructing the image is the decoder, which receives a 16×10 matrix from the digital capsule layer, reconstructs a 28×28 image after three full connection layers.

TABLE I. STRUCTURE COMPARISON OF CAPSULE NETWORK AND DEEP CAPSULE NETWORK

	Capsule network	Deep capsule network
Convolution layer	Conv1: 256*9*9	Conv1: 512*9*9 Conv2: 256*5*5
Primary Capsule	9*9	5*5
Digit Capsule	One time dynamic routing Three iterations	Twice dynamic routing The main route has three iterations, and the secondary route has three iterations
FC		

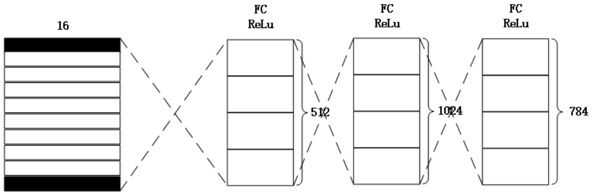


Figure 2. Decoder network

B. Working mechanism of dynamic routing

In this paper, there are two routes, the primary route and the secondary route, but both are the same dynamic path structure. It is used to ensure that the output of the capsule is only delivered to the appropriate parent node, which is similar to the idea of "focusing on Cultivation". It is necessary for the lower layer capsule i to know how to deliver its output vector to the higher-level capsule j. At this time, it is necessary to evaluate the coupling degree between the low-level capsules and the high-level capsules. This is represented by the scalar weight CIJ, which is the importance.

In this high-dimensional vector space, in order to describe the spatial relationship of different parts of the entity, each capsule is set with corresponding weight. An affine transformation matrix, which is composed of several weight vectors, an affine transformation matrix is generated. After transforming the matrix, we can get the **j** prediction vector $\hat{\boldsymbol{u}}_{\mathbf{j|i}}$ of each low-level capsule **I** to a high-level capsule. On the level of possible advanced capsules, the prediction vector $\hat{\boldsymbol{u}}_{j|i}=\mathbf{W}_{ij}\mathbf{u}_i$ is obtained by multiplying the weight matrix $\mathbf{W}_{ij}$ with the output $\mathbf{u}_i$ of low-level capsules. The prediction vector provides instance parameters for the capsule of high-level, and the higher-level capsule will be activated when the information provided by multiple prediction results is consistent.$\hat{\boldsymbol{u}}_{j|i}=\mathbf{W}_{ij}\mathbf{u}_i$

The low-level neural capsule **I** is connected with any high-level capsule **J** which has a "coupling" relationship with it. By multiplying the corresponding coupling coefficient $\mathbf{C}_{ij}$ with the **j** prediction vector $\hat{\boldsymbol{u}}_{j|i}$ of each low-level capsule **I** for the high-level capsule, and then weighted sum operation, the output $\mathbf{S}_j$ can be obtained. The output of the capsule in the next round is a high-dimensional vector $\mathbf{v}_j$, which is obtained by squash extrusion function on $\mathbf{S}_j$. The calculation formula is as follows:

$$S_j = \sum_i C_{ij} \cdot \hat{u}_{j|i} \tag{1}$$

For an intermediate layer capsule, the input is a vector and the output is also a vector, but the input process for it is two stages:

1) *Linear combination: For a linear combination of neurons, the connection weights between capsules are represented by a matrix instead of a scalar value in the form of a vector.*

2) *Dynamic routing:The core work of this stage is to determine the close relationship between high level **j** and low level **I**, that is to find the most suitable coupling coefficient value $\mathbf{C}_{ij}$, which is determined in the repeated process of dynamic routing algorithm.*

C. Dynamic routing algorithm

The process of dynamic routing algorithm is as follows:

a. Softmax processes data
b. Predict the output
c. Weighted sum
d. Compress the vector
e. Update coupling coefficient

The following figure 3 is a description of the dynamic routing algorithm.

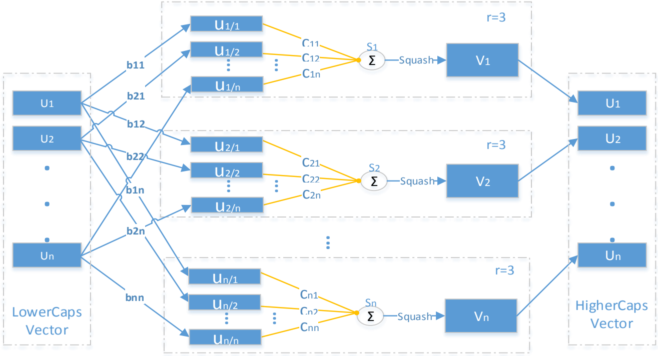


Figure 3. Dynamic routing algorithm

1) *The three input parameters are $\hat{\boldsymbol{u}}_{j|i}$ (prediction vector from **i** to **j**), **r** (number of iterations of routing algorithm) and **l** (number of layers of capsule).*

2) *For all layers of **l** capsules and (**l** + 1) capsules*

$$b_{ij} \leftarrow 0$$

For all layers of **l**capsules and (**l** + 1)capsules, the prior probability coefficient $\mathbf{b}_{ij}$ of two adjacent layers is initialized to 0, and its value will be used in the iterative update process of $\mathbf{b}_{ij}$. After the iteration, the value is stored in the corresponding $\mathbf{C}_{ij.}$

3) *Iteration with **r***

4) *The softmax rule is used to calculate the $\mathbf{C}_{ij}$ between the lower and higher layers.* In the beginning, because all $\mathbf{b}_{ij}$ are initialized to zero, so the obtained $\mathbf{C}_{ij}$ is also equal. That is to say, in this period, every node in the lower layer is equally important for the high-level capsule. The parent node at the higher level receives all the information from the lower level capsule. This kind of confusion in the initial stage of the algorithm will gradually become clearer in the later iterative calculation.

5) *The weighted sum of high-level capsules was calculated.* The weight of the combination used is the $\mathbf{C}_{ij}$ obtained in the previous step.

6) *$\mathbf{S}_j$ is a vector with a size and a direction.* However, if you want its length to be used as the probability of the existence of the entity, you need to normalize its size, and you need to use a nonlinear extrusion function to complete the normalization operation. This function retains the vector direction, and at the same time, the module length of the vector can be compressed within 1, so the output is the output of the high-level capsule

7) *The coupling between capsules is dynamic.* According to the formula, the larger the result value of $\hat{\boldsymbol{u}}_{j|i}{\cdot}\mathbf{V}_j$, that is, the more identical the pose information

of $\hat{\boldsymbol{u}}_{\mathbf{j|i}}$ and $\mathbf{V_j}$,the greater the value of $\mathbf{b_{ij}}$, which indicates that the coupling degree between the previous layer capsule **i** and the high-level capsule **j** is higher

Dynamic routing algorithm focuses on clustering similar parts together, and then forms a larger granularity identification module. If the predicted vector $\hat{\boldsymbol{u}}_{\mathbf{j|i}}$ and the output $\mathbf{V_j}$ of one of the high-level capsules are calculated by dot product and the result is very large, the relationship between nodes in this layer and high-level capsules will be strengthened after a reflection from front to back, that is to say, the coupling coefficient will be increased. At the same time, reduce the coupling coefficient with other high-level capsules. After **r** iterations, count the output of all high-level capsules, and determine the relevant routing parameters. The forward propagation will enter the next capsule layer of the capsule network.

D. loss function

The traditional cross entropy function only supports the scenario of one classification, so this function is not suitable for capsule network. In order to distinguish multiple classifications in a picture, the edge loss function is used to achieve the objective function of model optimization for each digital capsule **k**. It is shown in the following formula.

$$L_k = T_k max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k)max\left(0, \left||v_k|\right| - m^-\right)^2 \qquad (2)$$

In the above formula, $\mathbf{T_k}$ is the function of classification, **k** is the classification, the value of $\mathbf{T_k}$ is related to the existence of the k-th classification, $\mathbf{L_k}$ is the calculated loss. If and only if the **k** classification exists, $\mathbf{T_k}$ is 1; if there is no **k** classification, $\mathbf{T_k}$ is 0.|| $\boldsymbol{v_k}$ ||Represents the length of $\boldsymbol{v_k}$, which is the probability that the number **k** exists.m^+, m- are the threshold functions indicating the strength of the connection between the capsules .When it is lower than 0.1, it is considered that there is no connection relationship at all, and it is regarded as complete connection if it is higher than 0.9. In detail, $\boldsymbol{m^+}$ is the upper edge threshold, which is used to deal with the situation that the classification does not exist but exists in the prediction; $\boldsymbol{m^-}$ is the lower edge threshold, which deals with the situation that the classification does exist but is not predicted by the network.$\boldsymbol{\lambda}$ is called the sparsity coefficient and is used to adjust the weight between the two thresholds to adjust the parameters and steps. The values of $\boldsymbol{\lambda}$ are 0.9, 0.1 and 0.5. Add the loss $\mathbf{L_k}$ of each number to get the overall loss of the network.

IV. EXPERIMENT

A. Experimental environment

TABLE II. EXPERIMENTAL ENVIRONMENT

Operating system	Windows10(RAM16.0GB）
CPU	Intel(R)Core(TM)i7-9750H
GPU	NVIDIA GeForce GTX 1660 Ti
Dataset	MNIST
Other	pytorch1.5.0+cu101 python 3.7.7

B. Experimental data analysis

Handwritten digital machine vision database is widely used in image recognition and classification. The sample image in MNIST is 28 × 28 pixels, including four files: training set image, training set label, test set image and test set label. These files are binary files, each pixel of which is converted to a number between 0 and 255, where 0 is white and 255 is black. The training set has 60000 handwritten training samples. Its function is to fit model parameters, such as calculating offset and weight. The test set has 10000 samples, and its function is to test the final effect of the model.

1) Precision of capsule network test

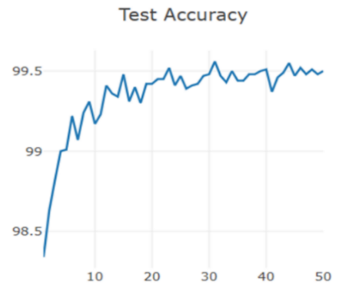


Figure 4. Test precision chart of capsule network under 50 epochs

As shown in Figure 4, the highest accuracy of this training is 99.55% in the 44th epoch.

2) Test precision of deep capsule network

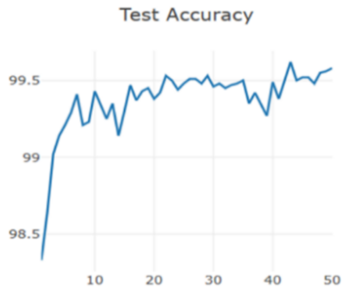


Figure 5. Test accuracy chart of deep capsule network under 50 epochs

As shown in Figure 5, the highest accuracy of this training is 99.62% in the 43rd epoch.

3) When epoch = 30, the final test accuracy of capsule network is 99.46%

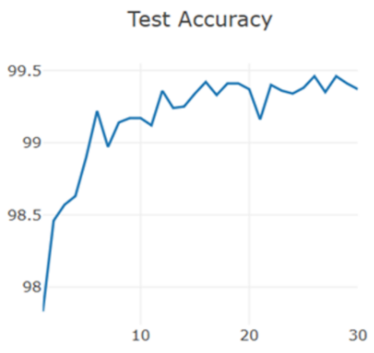


Figure 6. Test precision chart of capsule network under 30 epochs

4) When epoch = 30, the final test accuracy of deep capsule network is 99.58%

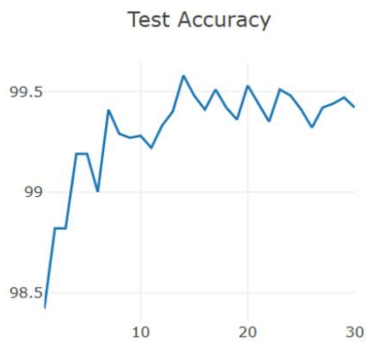


Figure 7. Test accuracy chart of deep capsule network under 30 epochs

5) Accuracy comparison between deep capsule network and capsule network

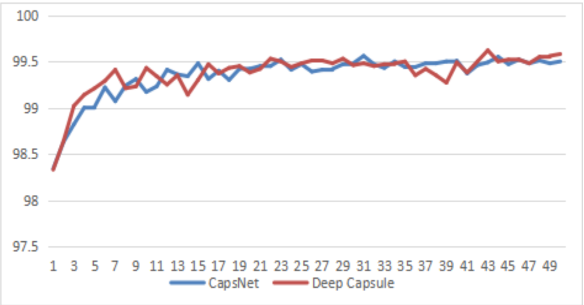


Figure 8. Comparison between the accuracy of capsule network and deep capsule network

As shown in Figure 8, it can be seen that the test accuracy of the deep capsule network in a short epoch increases faster than the accuracy of the capsule network and the recognition accuracy is also higher, under the same conditions.

6) Performance of deep capsule networks with the same two routes: 1,2,3,4

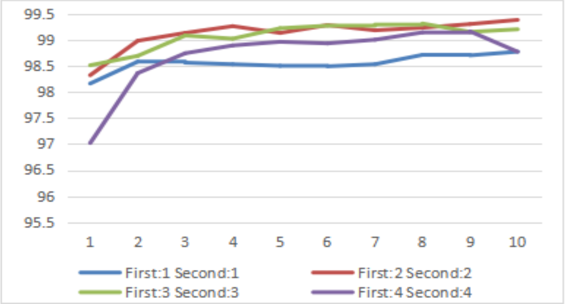


Figure 9. Impact of changing the number of routing iterations on the deep capsule network

As shown in the Figure 9, it shows that the number of route iterations is not the more the better, which should be obtained according to the specific experiment of network structure. In a smaller training period, It is more appropriate to select the number of iterations of the primary route 2 times and the number of iterations of the secondary route 3 times.

When the two routing iterations are different as to Figure 10.

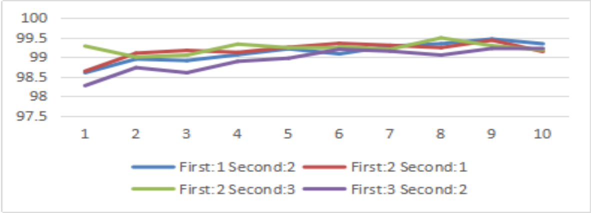


Figure 10. Influence of different iteration times of two routes on deep capsule network

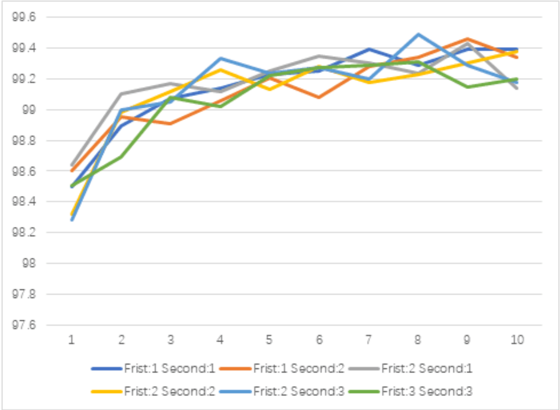


Figure 11. Influence of the same number of two routing iterations on deep capsule network

As shown in the Figure 11, the training time and classification accuracy of the network are compared under different collocation times of "primary route" and "secondary route". From the analysis of the data in the table, if only from the classification accuracy, the combination of "main route" iteration twice and "secondary route" iteration three times is the best, but the training time is long. If the training time and classification accuracy are considered comprehensively, the primary route is best to be iterated once and the secondary route is iterated twice.

C. Reconstruction

In order to understand the reconstructed picture, use the imshow function of matplotlib to draw and visualize, then the input picture 12 and the reconstructed picture 13 are shown in the following figure:

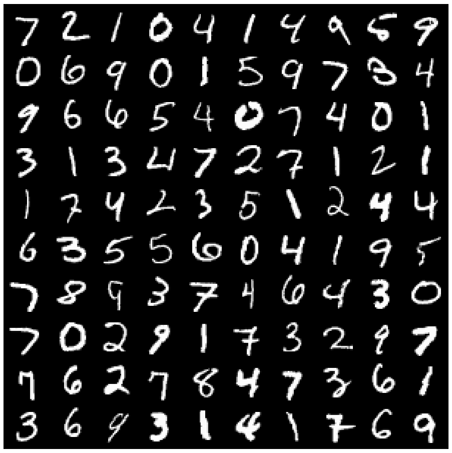


Figure 12. Schematic diagram of some pictures in MNIST database

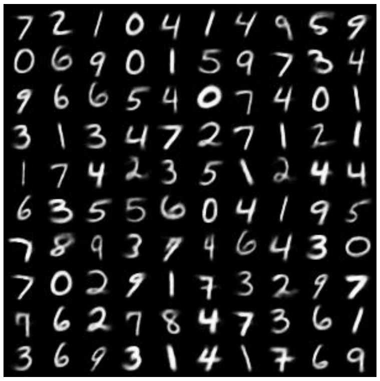
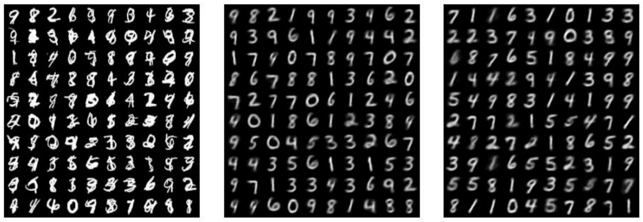


Figure 13. Schematic diagram of reconstructed image

From the comparison, we can see that the reconstructed digital image is clearer and smoother than the input image. It can be inferred that the reconstructed image has the function of smoothing noise.

D. Separation of overlapping handwritten numerals

In the same way, we train the overlapped handwritten digital images with deep capsule network, and finally put the vectors into the decoder to decode the reconstructed images. Some of them are shown in figure 14, and the separation effect is basically accurate.



Overlapping digital pictures

Reconstructed image set 1

Reconstructed image set 2

Figure 14. Comparison of input and output images of the network

Figure 15 shows the three separation effects' 0 'and' 1 ',' 3 'and' 4 ', ' 0 'and' 9 '. It is obvious that the network has been able to separate two completely coincident handwritten digits. Even if' 3 'and' 4 'overlap and it is difficult for human eyes to separate them, the network can still successfully separate them, with an accuracy rate of 93.53% .The accuracy rate of collaterals was only 88.10%.

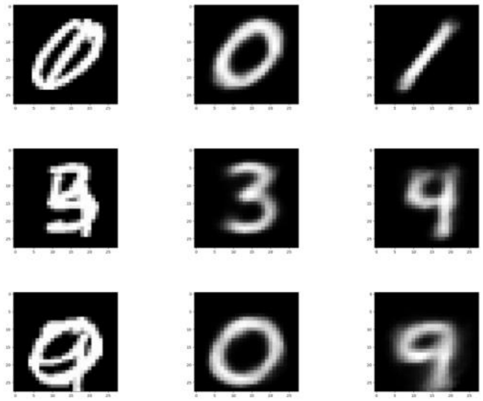


Figure 15. Partial reconstruction results of the improved network

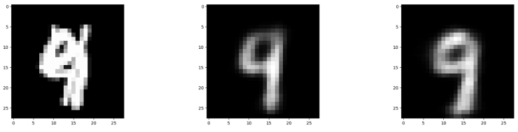


Figure 16. Improved partial error reconstruction results

However, the situation shown in Figure 16 still exists in the reconstructed picture. The original overlapping picture is the overlap of the numbers '9' and '4'. The two reconstructed images are like '9', without '4', the reconstruction is wrong, the error rate after the improvement is still 6.47%.

V. CONCLUSION

The deep capsule network model in this paper is based on the characteristics and shortcomings of the capsule network. On the one hand, it retains the advantages of capsule network in understanding the attitude of objects; on the other hand, in view of the shortcomings of capsule network, the convolution kernel size of convolution layer is optimized, and the dynamic routing process is improved to twice routing. The final deep capsule network not only retains the advantages of traditional capsule network, but also improves the performance.

REFERENCES

[1] Hinton G E, Krizhevsky A, Wang S D. Transforming auto-encoders[C]//International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2011: 44-51.

[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[3] Hinton, Geoffrey E.; SABOUR, Sara; FROSST, Nicholas. Matrix capsules with EM routing. 2018.

[4] Kosiorek A, Sabour S, Teh Y W, et al. Stacked capsule autoencoders[C]//Advances in Neural Information Processing Systems. 2019: 15512-15522.

[5] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[6] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[7] LeCun, Yann, Corinna Cortes, and Christopher JC Burges. "The MNIST database of handwritten digits, 1998." URL http://yann. lecun. com/exdb/mnist 10 (1998): 34.

[8] Netzer, Yuval, et al. "Reading digits in natural images with unsupervised feature learning." (2011).

[9] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533-536.

[10] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[C]//Advances in neural information processing systems. 2017: 3856-3866.

[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[12] Zhao W, Ye J, Yang M, et al. Investigating CapsuleNetworks with Dynamic Routing for Text Classification[J].arXiv preprint arXiv:1804.00538, 2018

Improve Usability of Tourism Websites Based on Agile Strategies

Tsegaye Yabsra Asefa
School of Computer Science and Engineering
Xi'an Technological University
E-mail: yabsraasefa@gmail.com

Abstract—**Usability is critically important in developing tourism websites because visitors expect tourism websites to be attractive, interactive and informative. The main aim of this paper is to analyze and examine the content of tourism websites. A case study has been performed on a tourism website to understand the usability gaps of tourism websites development process for better website usability and user satisfaction. The paper discusses factors that are success combinations of features that a tourism websites must have. Considering the success factors10 tourism websites were analyzed. The paper presented a report on the analyzed data and give suggestions for making tourism websites more efficient.**

Keywords-Usability; Agile Strategy; Content Analysis; Tourism Websites; Progressive Web Apps; Predictive Search

I. INTRODUCTION

The wide variety of existing information and the rapid development of information technology have made many web-based application products and services available for dally use. Information distribution can be done by improving knowledge access and transferring knowledge by using media such as website [2]. Websites are used daily for reading news, finding work vacancy, shopping, finding telephone information, ordering food, planning for a trip, selling products and even helping a company business processes. Among such website is tourism website. Interface development for tourism website must actively involve user from planning trough evaluation [3].

On the Web, usability is a necessary condition for survival. If a website is difficult to use, people leave [6]. According to Hayat Dino [7] usability must be tightly integrated in the software development process and makes its objective to examine how usability can be integrated more tightly into a specific software development process model that is being applied by the majority of software development companies. The lack of usability may lead the end-user to need further assistant and having bad perception about the application or the website in essence that the end-user becomes less satisfied to use the application.

As Hayat Dino mentioned usability must be tightly integrated in the software development process and makes its objective to examine how usability can be integrated more tightly into a specific software development process model that is being applied by the majority of software development companies. The lack of usability may lead the end-user to need further assistant and having bad perception about the application or the website in essence that the end-user becomes less satisfied to use the application.

An agile organization is defined as one that: adapts its organizational culture to market change, learns about market changes rapidly, takes advantage of these market changes and customizes its products to individual preferences (Desouza, 2007). An agile organization has the ability to take advantage of its opportunities and deploys its tangible and intangible assets in quick time cycles with the lowest possible cost and effort.

For Salah, Paige and Cairns [36], Agile and usability Integration gained increased interest due to these reasons:

- The reported advantages of Usability on the developed software as it enables developers to understand the needs of the potential users of their software, and how their goals and activities can be best supported by the software thus leading to improved usability and user satisfaction.

DOI: 10.21307/ijanmc-2020-032

- The Agile community hardly discusses user needs and user interface design. More-over, none of the major Agile processes explicitly include guidance for how to develop usable software.

A.Nilawati,D.R.A.Pratama et al. [3] conducted analysis on Indonesia official tourism website using criteria specified for usability testing such as effectiveness, efficiency, consistency and interface design. Usability testing measures related to the human computer interaction used including ease of use, easy to learn, errors and syntax. As a conclusion the researchers suggested that a web-page design must be user oriented and the importance of identifying what kind of user that will visit the website. Even if the researchers evaluated the website based on evaluation criteria and identify usability gaps, they didn't address the website development methodology used with respect to website usability.

This paper performs a case a study on a tourism website to understand the usability gaps of tourism websites development process for better website usability and user satisfaction. Section two presents methods used to analyze tourism website Section 3 discuss success factors that a tourism website used have. Section 4 gives suggestions for the coming researchers Section 5 reports the outcome.

Present paper discusses several issues and challenges related to web usability. Analysis of various web usability factors is also presented. Rest of paper is organized as: Section 2 presents web usability and its significance. Section 3 discusses numerous issues and challenges related to web usability. In section 4, the proposed solution about good and usable web design is presented. Section 5 describes the conclusion of the paper.

II. METHODOLOGY

A case study will be performed on a tourism website to understand the usability gaps of tourism websites development process for better website usability and user satisfaction.

In order to have a thorough view of the usability of current city tourism websites, this study examines 10 sample city tourism websites. These 10 tourism websites were selected from https://blog.feedspot.com/africa_travel_blogs

In order to examine the content of the website I used content analysis methodology.

As Amy Luo definition

Content analysis *is "a research method used to identify patterns in recorded communication. To conduct content analysis, you systematically collect data from a set of texts, which can be written, oral, or visual, Books, newspapers and magazines, Speeches and interviews, Web content and social media posts and Photographs and films*

Hence, content analysis can be an effective tool to examine the content and functions of city tourism websites. The homepage of each of the 10 websites was thoroughly examined and the data were recorded in terms of functions provided, services offered and content presented. The results were grouped in several ways: by nations, by attributes, etc.

A. Selected tourism websites

1) Getaway Travel Magazine (GTM)

2) Jumia Travel Blog(JTB)

3) Cape Town Tourism(CTT)

4) In Africa and Beyond(IAaB)

5) Iconic Africa(IA)

6) Jumia Travel(JT)

7) Make my Trip(MmT)

8) Wego(W)

9) Africa's Jewel(AJ)

10) Nomadic Holidays and Safaris(NHS)

III. FEATURES OF TOURISM WEBSITES

These days, travelers use online travel portal to meet all their travel needs. People want to plan their vacations before they go, and they can't do that if your site isn't user-friendly

First impressions are formed quickly which also applies to travel websites now more than ever. Time spent by new users on the website is limited which means you have less time to get their attention and keep them in your website. [25] For travel website features, images and a standard booking system won't suffice. Over the years the web has changed drastically and something engaging and exciting is required to strike a chord with customers.

With this situation, the following factors are some of the winning combinations of features that a travel websites must have.

➢ Social sharing buttons

➢ Sticky menu & back to top button

- Augmented Reality (AR) and Virtual Reality(VR)
- Progressive Web Apps (PWAs)
- Images
- Predictive Search
- Pricing
- Easy Booking System

IV. CASE STUDY OF CITY TOURISM WEBSITES

Considering the above tourism website success factors the 10 tourism websites were analyzed.

A. *Social Sharing Buttons*

The key to getting someone to do something is to make it easy for them. Adding social sharing buttons means that visitors can share the site to their own social media pages with one click. [25]

- It incorporates the number of times content is liked, shared, viewed, etc. So the more people you have sharing, the better you look to Google and the higher you are in the search results.
- Friend groups and families tend to travel together or to similar places. When the time comes to book a vacation, people are going to check what their friends recommend.

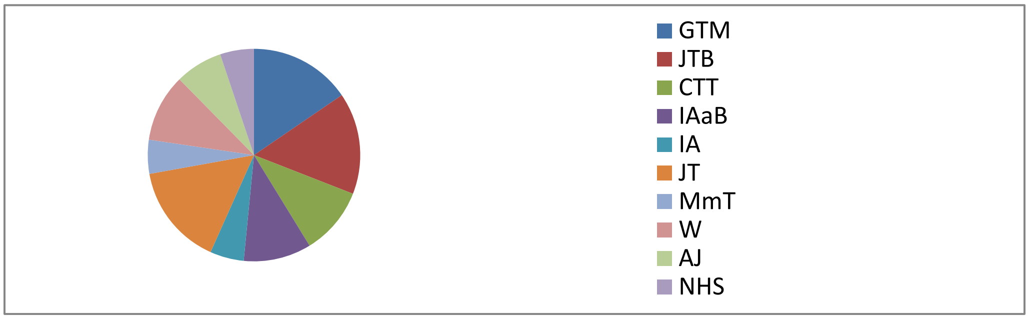


Figure 1. Persentage of website having social sharing butttons

B. *Sticky menu & back to top button*

With a sticky menu or back to top button, your tourism website visitors will be able to quickly get to where visitors want to be.

TABLE I. PERSENTAGE OF WEBSITE SHOWING STICKY MENU AND BACK TO TOP BUTTON

GTM	JTB	CTT	IAaB	IA	JT	MmT	W	AJ	NHS
90%	85%	75%	80%	65%	90%	80%	90%	80%	80%

C. *Augmented Reality (AR) and Virtual Reality(VR)*

AR provides a virtual tour with a 3D view of reviews for nearby location, Wi-Fi hotspots, real-time weather forecast and more. It can show extra information like destination information, eating joints and more. [26] It allows hotels and other businesses in this field to enhance the physical environments like local sights and hotel rooms so that customers will be encouraged to visit the place.

Virtual gives guided tours of any place in the world. This will especially help travelers explore small and less-known places. By giving a 360-degree view of the different locations, travel companies and agents can let customers explore the ground before booking and increase the level of trust simultaneously. [26]

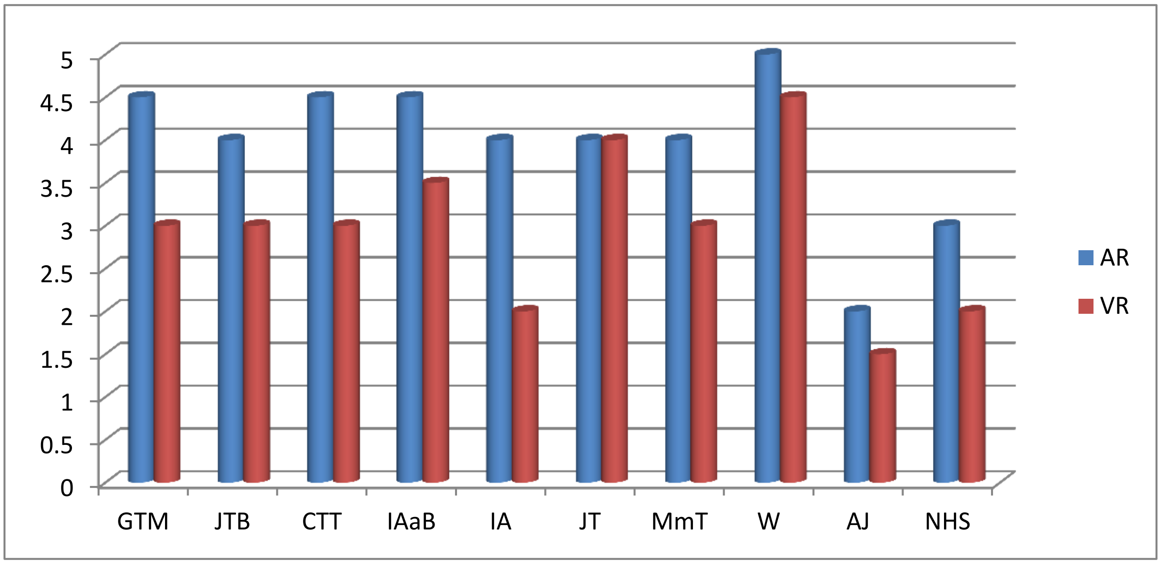


Figure 2. Persentage of website showing Agumented reality and Vertual reality

D. Progressive Web Apps (PWAs)

Progressive web apps are modern web pages that also act as mobile apps. These apps have the usefulness of a native app but, when accessed through a browser, it does not require any downloading of apps which is a huge plus on conversion and usage ratio. If a user books a hotel through PWA, the user can access the information via the browser without internet connectivity too. Additionally, the PWA web page can be saved on the user's home screen and used as a mobile app. [26]

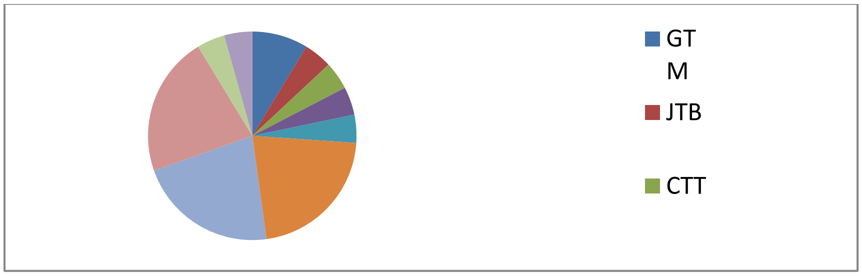


Figure 3. Persentage of website having Progresive web Apps

E. Images

We live in a visual world and images are the most powerful way to inspire and transmit messages. They make a strong statement and will have a bigger impact on your website. Also, people interact more in social networking sites that has plenty of images. So it's safe to say that users will stay longer and interact more in a website that has lovely images.

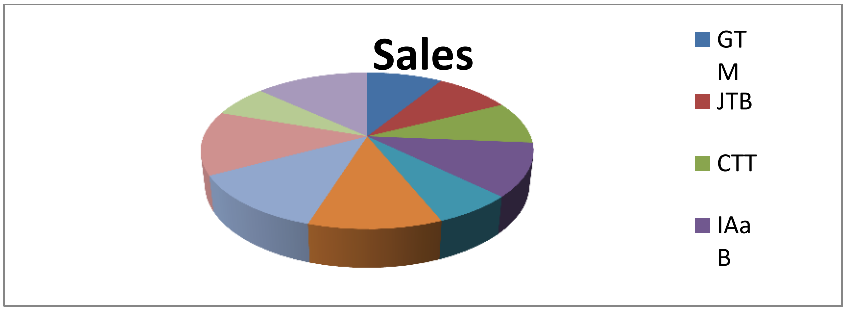


Figure 4. Persentage of website showing image

F. Predictive Search

Predictive search is nothing but a drop-down list that pops up immediately while typing, so that you don't have to hit 'search' to find out common queries. This is an excellent way to avoid displaying a large list of results, finding results quickly and to display questions the user hasn't thought of. It saves on typing if the user is using a device which does not have a physical keyboard like a tablet. [26]

TABLE II. PERSENTAGE OF WEBSITE HAVING PREDICTIVE SEARCH

GTM	JTB	CTT	IAaB	IA	JT	MmT	W	AJ	NHS
80%	75%	80%	85%	95%	80%	75%	90%	80%	75%

G. Pricing

If the pricing is not clear and has hidden charges and taxes, there is a high chance of the user spending their money elsewhere. So, if you want to increase your profits, it is essential to know the importance of clear pricing. A pricing table must be simple and clear so that users' can choose the appropriate package. [26]

TABLE III. PERSENTAGE OF WEBSITE SHOWING PRICE

GTM	JTB	CTT	IAaB	IA	JT	MmT	W	AJ	NHS
80%	0	0	0	0	1	1	1	0	0

H. Easy Booking System

To provide real-time updates, make sure your website is linked to a property management system (PMS). It will make things easier as they have to provide information like destination, check-in and check-out dates, contact details, and other travel related data [26]

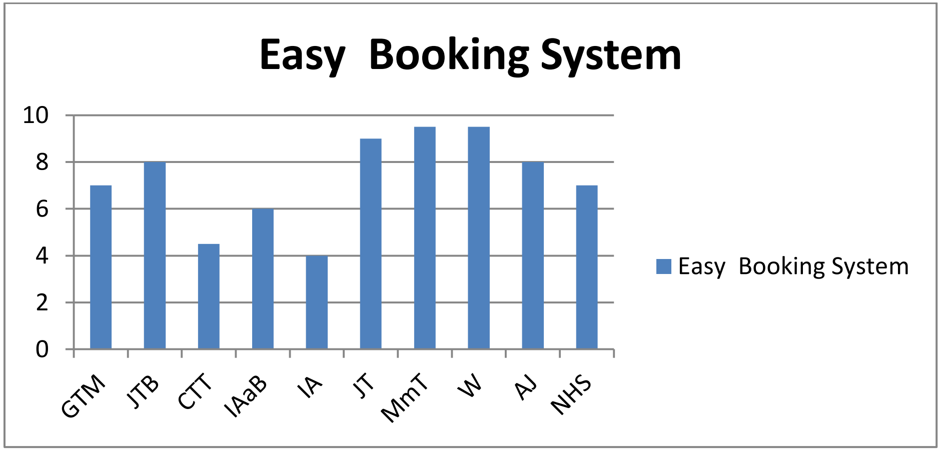


Figure 5. Persentage of website having Easy booking system

V. SUGGESTIONS

I suggest the following solutions to improve the usability of city tourism websites and to bridge the gap among visitors.

➢ Adding Social sharing buttons

➢ Augmented Reality (AR) and Virtual Reality(VR)

➢ Presenting information for specific users would also help them easily find the information they need and build trust between the website and the users

➢ Broadening the way to engage the users and Progressive Web Apps (PWAs)

➢ Encouraging Image exhibition/gallery especially increasing interactivity Pricing

➢ Adding easy Booking System

VI. CONCLUSION

This research takes a close look with content analysis at representative tourism websites across http://www.umsl.edu/~wilmarthp/mrpc-web-resources/content-analysis.pdf. Usability here plays a crucial role as it is how easy is to use the site. It shows that usability challenges exist in various aspects in both content and functions. From this research I have learnt that we should create our site keeping in mind all the above factors given but check before updating it.

Many problems reflect a lack of user-centered design and interactivity in the city tourism website development. Also, I found that a site becomes irritating and user leaves the site in just 2 sec when the users demand or aspects are not fulfilled. So this, the most important is to make your first impression very calm and worth, so that the user can at least get some interest to know about the site. A well designed tourism website should consider the distinct navigating and information seeking behaviors of different users which is especially vital in the international tourism industry.

VII. LIMITATION AND FUTURE RESEARCH

I wish I had more time in doing this research. As we know one limitation lies in a pandemic disease so I'm expected to stay home with that it's a bit difficult to go through, understand and come up with an updated and new report. I also had a limited number of sample websites. Future research can be done in studying multiple tourism website

REFERENCE

[1] D. Murugaiyan, "Waterfall Vs V-Model Vs Agile: A Comparative Study on SDLC," Int. J.Inf. Bus. Manag., vol. 2, no. 1, pp. 26–30, 2012.

[2] D. Lon, "Importance of DMO Websites in Tourist Destination Selection," pp. 373–385,2013.

[3] A. R. Nilawati, D. A. R., A. Y. Pratama, D. Adlina, and N. R. Al Mukarrohmah, "Interface on usability testing Indonesia official tourism website," Int. J. Hum. Comput. Interact., vol. 3, no. 2, pp. 26–34, 2012.

[4] Buhalis, D., and Law, R. (2008) Twenty years on and 10 years after the Internet: The state of eTourism research. Availabe at: http://eprints.bournemouth.ac.uk/5126/1/TMA_eTourism_20years_Buhalis%26Law_FI NAL_.pdf, Accessed May 02, 2015.

[5] Ali Fathi Khomeyrany, "Agile strategies in eTourism" 2015

[6] GOOGLE, "2014 Traveler Road To Decision," no. June, 2014.

[7] R. BAGGIO, "A Websites Analysis of European Tourism Organizations," no. May, pp. 1–15, 2014.

[8] Jacob Nielsen, "Usability 101: Introduction to Usability," 2012.

[9] H. D. BEDRU, "A Framework for Integrating Software Usability Into SoftwareDevelopment," no. June, 2012.

[10] Levén, P., and Holmström, J. (2012) Regional IT innovation: a living lab approach. International Journal of Innovation and Regional Development, 4(2), pp. 129-143

[11] Lin, Y., Poschen, M., Procter, R., Goble, C., Bhagat, J., & Roure, D. D. (2008) Agile Management: Strategies for Developing a Social Networking Site for Scientists. University of Manchester. Availabe at: http://eprints.soton.ac.uk/265854/1/eSS_myEx_final.pdf , Accessed May 02, 2015.

[12] Pressman, R. (2009) Agile Development. academic.brooklyn. Availabe at: http://academic.brooklyn.cuny.edu/cis/sfleisher/Chapter_03_sim.pdf, Accessed May 02, 2015.

[13] Rönnbäck, L., Holmström, J., & Hanseth, O. (2007) IT-Adaptation Challenges in the Process Industry: An Exploratory Case Study. Industrial Management & Data Systems, 107(9), pp. 1276-1289

[14] N. Vatankhah, K. T. Wei, and S. Letchmunan, "Usability measurement of Malaysian online tourism websites," Int. J. Softw. Eng. its Appl., vol. 8, no. 12, pp. 1–18, 2014.

[15] A. F. Khomeyrany, "Agile Strategies in eTourism A case study of Umeå Tourism Information Centre," 2015.

[16] K. C.DESOUZA, Agile Information Systems Conceptualization, Construction, and Management, vol. 39, no. 5. 2008

[17] R. S. Pressman, CS605-Software Engineering Practitioner's Approach. 2010.

[18] Rönnbäck, L., Holmström, J., & Hanseth, O. (2007) IT-Adaptation Challenges in the Process Industry: An Exploratory Case Study. Industrial Management & Data Systems, 107(9), pp. 1276-1289

[19] Westergren, U., and Holmström, J. (2012) Exploring preconditions for open innovation: Value networks in industrial firms. Information and Organization, 22, pp. 283-294.

[20] Z. Kribel, "Information and Communication Technology in Tourism," Inf. Commun.Technol. Tour. 2009, pp. 123–134, 2009.

[21] J. Venkatesh, "Improving User Experience by using Agile Methodologies," pp. 14–17, 2012.

[22] Amy Luo, Dan, "What is content analysis and how can you use it in your research?" July 18, 2019

[23] https://www.tourismtiger.com/blog/5-tourism-website-must-haves/

[24] https://colorwhistle.com/travel-website-features/

Health Assessment Based on D-S Evidence Theory of Equipment

Tanghui Sun
School of Computer Science and Engineering
Xi’an Technological University
Xi’an, China
E-mail: 1017540991@qq.com

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

***Abstract*—Prognosis and Health Management (PHM) is the important technical means to achieve condition based maintenance(CBM), and Health assessment is the important part of the PHM system. PHM is widely used in aerospace area, But it lacks application in ground field, In order to improve the accuracy of self-propelled gun system health assessment, taking a certain self-propelled gun system as the object, in this paper, takes self-propelled gun as an example, divide it’s health status into 5 levels, for equipment that fails the test, it can be directly determined to be in a fault state; for equipment that has passed the test, and come up with a model of health assessment that based on D-S evidence theory, first, Select the health indicators which can comprehensively represent the status of propelled artillery, and use the normalized quantization method to Process the data. Then, employ the D-S evidence theory to make the integrated decision on the membership of each health parameter after treatment, finally, determine the ultimately health status of self-propelled artillery, the rationality of the method is verified by experiments.**

Keywords-Self-Propelled Guns; D-S Evidence Theory; Health Assessment

I. INTRODUCTION

At present, the elimination of equipment’s abnormal state mainly rely on regular maintenance and break down maintenance, and that exposes many disadvantages in the modern equipment maintenance support activities. Under such circumstances, the need for condition based maintenance(CBM) is increasing. Under this background, PHM was born at the right moment. Health assessment is an important function of equipped PHM system, Make correct assessment of the health status of equipment, not only provide a basis for equipment fault prediction and maintenance decision, and provide technical support for the accuracy of equipment maintenance, and it is of great significance to performance, operational and environment of improve the combat effectiveness and service life of artillery.in this paper, take self-propelled guns as the research object to carry out health assessment research. From the perspective of health status assessment, first, Select the health indicators which can comprehensively represent the status of propelled artillery, and use the normalized quantization method to Process the data. Then employ the D-S evidence theory to make the integrated decision on the membership of each health parameter after treatment, finally, determine the ultimately health status of self-propelled artillery.

DOI: 10.21307/ijanmc-2020-033

II. HEALTH ASSESSMENT SYSTEM AND HEALTH CLASSIFICATION

A. Analysis of health indicators

To evaluate the health status of self-propelled artillery, the health status parameters of each system should be determined. In order to evaluate the health of self-propelled gun effectively, it is necessary to determine the health parameters of each self-propelled gun system. The following factors should be taken into account when selecting the parameters of self-propelled artillery's health state assessment:

1) Selected parameters can effectively represent the health state of the equipment;

2) The selected parameters should be convenient for collection and mutually independent;

3) Consider the use factor.

There are many test items for self-propelled artillery, and different items represent different performance. To establish a comprehensive and reasonable health index system is the first problem to be solved for self-propelled artillery health assessment. The establishment of a health indicator system should consider both comprehensiveness and incompatibility. The performance testing content of self-propelled gun mainly includes performance, maneuverability and environment. Self-propelled gun is a complex system, and the environment is harsh, under the existing conditions, it is impossible to measure all the indicators. To analyze and evaluate the health status of artillery, the key factors should be extracted from the perspective of reflecting the artillery, rather than using all indicators.

Figure 1 shows the health evaluation system of self-propelled artillery.

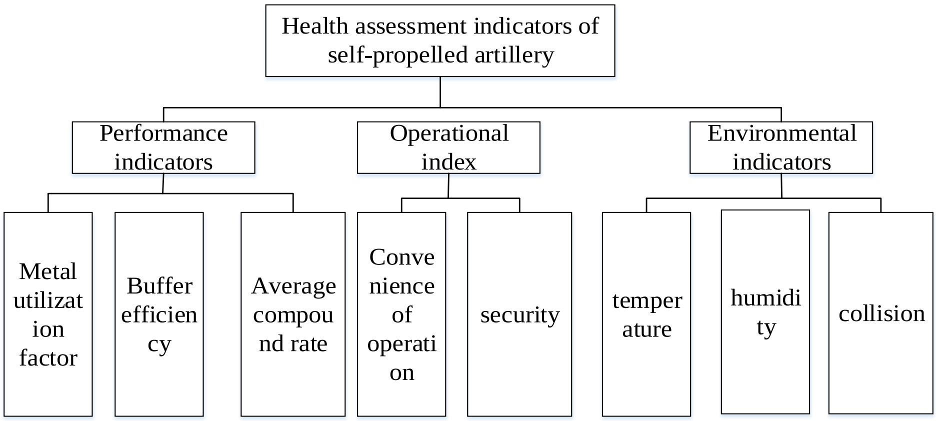


Figure 1. Health assessment indicators

The health indicators of self-propelled guns are divided into three parts: performance indicators, operation indicators, and environmental indicators. In terms of performance indicators, in order to facilitate the quantification, from the perspective of reflecting the performance of the artillery, three factors that can reflect the quality and health of the artillery are extracted to evaluate the overall health of the artillery. Its set is u=(u1,u2,u3), where u1 is the metal utilization coefficient; u2 is the buffer efficiency; u3 is the average Re-advance rate. The muzzle momentum metal utilization factor refers to the muzzle momentum provided by the mass of the artillery, which is a comprehensive consideration of the power and mobility of the artillery; the buffer efficiency and the average return speed are used to measure the launch speed of the artillery.

The environmental information is mainly the environmental information of the operating scene of the self-propelled artillery system. The operating scene of the self-propelled artillery system is largely restricted by the environment, and its performance is also affected by environmental factors, thereby affecting its health level. Therefore, environmental information is also an indispensable data for evaluating the operating status of the system; The operation mode of modern self-propelled artillery weapon system has changed from manual to semi-automatic and fully automatic, and the driving mode has also changed from towed to self-propelled. The interaction between the gunner and the artillery is more diversified, and the space for movement is further restricted. The relationship between environment and environment is more complicated. This article starts from the operational requirements of "convenience, safety, and accuracy", and focuses on examining the impact of artillery and the environment on personnel. Based on this, the main factors affecting the operability of the artillery are maneuverability factors and safety factors. Maneuverability factors refer to factors that directly affect the completion quality of operations or reduce operational efficiency, and safety factors refer to factors that affect the safe use of artillery.

B. *Health status classification*

In the past, when evaluating the health status of self-propelled artillery, the “right-and-no” system was often used, that is, the health status of self-propelled guns was simply divided into qualified and unqualified. It is reasonable that the test data falls within the specified threshold range, and it is not to exceed the specified threshold. This “yes or no” assessment method may use the same maintenance strategy for equipment that is in very good condition and equipment that is close to failure. This will cause unnecessary repairs for the former, and may cause insufficient repairs for the latter. It affects its combat readiness and cannot achieve condition based maintenance of equipment. Therefore, it is considered to refine the level of equipment health. However, the classification of health status levels should not be too many, otherwise it may not be possible to determine which maintenance measures to take for equipment with different health status levels. According to the requirements, the status of self-propelled artillery (or indicators) can be divided into 5 status levels of health, good, attention, deterioration and failure, as shown in Table 1 below.

TABLE I HEALTH STATUS LEVELS AND DESCRIPTION

Health grade	grade description
health	The measured data are in the range and close to the standard parameter values, don’t require maintenance.
good	The measured data are all in the range, and some data are wandering in a small range, but far from the attention value, maintained as planned, Scheduled maintenance.
attention	The measured data are all within the range, and most of them swim and fail to reach the attention value.
deterioration	The measured data are all within the range, but some data are close to the attention value, which needs to be monitored and repaired as soon as possible.
failure	Some data have exceeded the value of attention and must be repaired and secured immediately.

According to the above definition, it can be considered that the state of health and good state belong to health, and the state of attention and deteriorating state belong to sub-health. The self-propelled artillery in "healthy" and "sub-healthy" states is qualified because the test data of all parameters are within the allowable range. However, for "sub-healthy" self-propelled artillery, it is necessary to attract the attention of maintenance personnel. In a certain period of time in the future, the self-propelled artillery in this state is likely to degenerate into a malfunctioning state, so monitoring must be strengthened. For a self-propelled artillery in a malfunctioning state, because the parameter test data exceeds the threshold, it is unqualified. In order to ensure its combat readiness and mission success, reasonable maintenance measures must be arranged immediately.

It should be mentioned that the state parameter of the system index does not necessarily indicate the health status level, and there is not necessarily a clear boundary between the two. It is necessary to normalize the membership function to transform the state parameter into the state level.

III. Health status assessment based on D-S evidence theory

Since the health status of self-propelled artillery is represented by the health status of multiple parameters, in order to determine the health status of self-propelled artillery, it is necessary to evaluate the health status of its parameters and determine the health status of each parameter. According to the health status classification of self-propelled gun, when evaluating the health status of parameters of self-propelled gun, we should first judge whether the parameters are out of tolerance according to the test results of parameters. If the test results of parameters exceed the threshold value, it indicates that the parameters are unqualified, then the self-propelled gun can be directly judged to be in fault state. Otherwise, it shows that the parameters are qualified and need to be further analyzed. The following is a health assessment of the parameters that pass the test. Without explanation, the parameters in this paper refer to the parameters that pass the test.

A. D-S Evidence Theory

D-S evidence theory aims at the results (evidence) after the occurrence of an event and explores the main causes (hypothesis) of the occurrence of an event. D-S evidence theory is an effective method to integrate subjective uncertain information for multi-attribute decision problems with subjective uncertain judgment. The D-S evidence theory was proposed by Dempster in 1967 and was further developed and perfected by his student Shafer in 1976, so it is also called the D-S evidence theory. D-S evidence theory is currently become one of the important tools for processing uncertain information and fusing multiple inference results. It can comprehensively consider the weight of each information in multi-source information and reduce the divergence of conclusions caused by multi-source information inference. The basic theory and methods will be introduced below.

1) Establishment of identification framework

In D-S theory of evidence, propositions are generally represented by sets, that is, questions that need to be decided, collection of all possible answers said with Θ ,the set Θ is called recognition framework, and can be represented as: $\Theta = \{\theta_1, \theta_2, ...\theta_n\}$, that is to say, by limited element of n (these elements are independent of each other and mutually exclusive) constitute a non-empty set Θ for recognition framework, can be used to represent all the possibilities of events, and called 2^{Θ} Proposition A can be expressed as A subset of Θ , namely $A \subseteq \Theta$, or $A \in 2^{\Theta}$. For each subset Θ can assign a probability, known as the basic probability distribution. For example, a flashlight can emit light of A, B and C, then the identification frame is $\Theta = \{A, B, C\}$, and result is $2^{\Theta} = \{\phi, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}$. And in this paper, the health status of self-propelled artillery is divided into five levels, health, good, attention, deterioration and failure.

2) Basic trust allocation function

The recognition framework contains N elements, and the basic trust allocation $m(A)$ of some evidence on this recognition framework. Is a collection of from 2^{Θ} mapping to $[0,1]$. For any $A \subseteq \Theta$, such as function $m(A) \to [0,1]$ satisfy conditions:

$$\begin{cases} m(A) = 0, A = \phi \\ \sum_{A \subseteq \Theta} m(A) = 1, A \neq \phi \end{cases} \quad (1)$$

Then called $m(A)$ is basic trust distribution function in recognition framework Θ. for any subset of A in Θ Framework, if $m(A) \succ 0$, called A as focal element. $m(A)$ indicates the degree of trust in A by the evidence. $m(\phi) = 0$, reflecting the fact that non credibility in empty sets (empty propositions).

When $A \subseteq \Theta$, and A consists of single element, $m(A)$ is the appropriate precise trust to A propositions; When $A \subseteq \Theta$, $A \neq \Theta$, and A is composed of multiple elements, $m(A)$ is precise trust to A propositions, is also the corresponding proposition but this part of which of the trust don't know should be assigned to the specific element of A; When $A = \Theta$, then $m(A)$ is the rest portion after each subset of Θ to do the trust assignment , and don't know how to allocate to it.

3) The trust function and likelihood function

There is the identification framework $\Theta, m(A)$ is the basic probability assignment on Θ, and there is a mapping $Bel: 2^{\Theta} \to [0,1]$ and $Pl: 2^{\Theta} \to [0,1]$, satisfying :

$$\begin{cases} Bel(A) = \sum_{B \in A} m(B) \\ Pl(A) = \sum_{B \cap A \neq \Phi} m(B) \end{cases} \quad (2)$$

The $Bel(.)$ and $Pl(.)$ functions in the formula, the former is used to represent the function trust function, and the latter is used to represent the likelihood function. The likelihood function $Bel(.)$ and the trust function $Pl(.)$ represent the upper and lower limits of the trust level of A primitive, Specifically, the following can be used to characterize the relationship between the two: $Pl(A) \geq Bel(A), A \subseteq \Omega$. In this way, When measuring the uncertainty of A, $[Bel(A), Pl(A)]$ Can be introduced, And on this basis, Define the probability of event A as $P(A) \in [Bel(A), Pl(A)]$. And the relationship between likelihood function and trust function is shown in Figure 2 below.

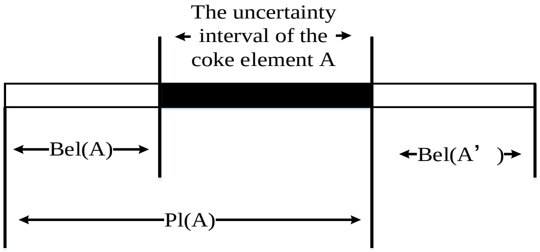


Figure 2. Trust function and likelihood function relationship

4) The synthesis rules of D-S evidence theory.

According to the definition of evidence theory, for the evaluating of framework for Θ, different features can obtain evidence of the different body probability distribution value $\mathrm{m}_1, m_2, \ldots\ldots, m_n$, then the combined process of the probability distribution of each evidence body is as follows:

$$\mathrm{m}(A) = \begin{cases} \frac{1}{1-K} \sum_{\cap A_i = A} \prod_{1 \leq i \leq N} m_i(A), A \neq \Phi \\ 0, A = \Phi \end{cases} \quad (3)$$

In the formula, $K = \sum_{\cap A_1 = \Phi} \prod_{1 \leq i \leq N} \mathrm{m}_i(A_i)$.

The D-S evidence theory synthesis principle can be summarized as $m(A) = m_1 \oplus m_2$. For the combination of multiple evidence, $M = m_1 \oplus m_2 \oplus ... \oplus m_n$, that can be generalized from the combination of the two evidences. And the K in the formula is used to represent the conflict between various evidences. K is 0 is a consistent evidence, used to indicate that there is no conflict. If the value is close to 1, the greater the conflict.

5) Decision of health status level

After the improved evidence theory is applied to synthesize the health status of multiple equipment parameters, in order to determine the final health status of the equipment, the decision method based on the basic probability assignment can be used to make a decision on the resultant health status of multiple equipment parameters, namely, the principle of maximum attribution.

Set $\exists A_1, A_2 \subset U$ as two health status levels of self-propelled artillery, which can be satisfied

$$\mathrm{m}(A_1) = \max\{\mathrm{m}(A_1), A_i \subset U\}, \tag{4}$$

$$m(A_2) = \max\{m(A_i), A_i \subset U and A_i \neq A_1\}, \tag{5}$$

For the pre-set threshold ε_1 and ε_2 , if

$$\begin{cases} m(A_1) - m(A_2) \succ \varepsilon_1 \\ m(U) \prec \varepsilon_2 \\ m(A_1) \succ m(U) \end{cases} \tag{6}$$

If so, the probability of A_1 is much higher than A_2, the final health state of self-propelled gun can be determined by this method. That is, the final health status of self-propelled artillery is A_1.

B. Steps based on D-S evidence fusion

There is an uncertain corresponding relationship between the health status of self-propelled artillery and its performance, operation, environment and other factors, which is fuzzy and unknown. Therefore, D-S evidence theory is introduced to carry out data fusion in the health assessment of self-propelled artillery. In the health assessment of self-propelled artillery, several indicators will produce a certain health state, and each indicator of health state has a certain probability of occurrence. In the D-S evidence theory, the probability is represented by the basic credibility distribution, and the different locations of the transmitter are tested by multiple sensors to obtain the basic credibility distribution that the measured indicators of each sensor belong to various health states. Then the D-S combination rule is used for information fusion to obtain the merged health assessment indicators.

The steps of health state assessment of self-propelled artillery based on evidence theory are shown in Figure 3.

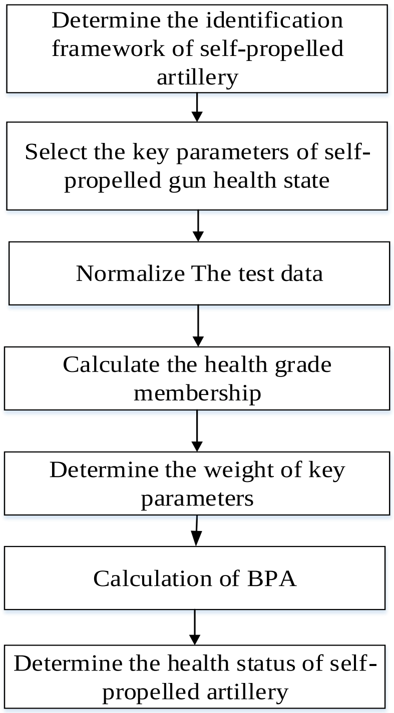

Determine the identification framework of self-propelled artillery
Select the key parameters of self-propelled gun health state
Normalize The test data
Calculate the health grade membership
Determine the weight of key parameters
Calculation of BPA
Determine the health status of self-propelled artillery

Figure 3. Health assessment procedures for equipment

As can be seen from the figure above, the health status assessment steps of self-propelled artillery are as follows:

1) Determine the identification framework

Determine evaluation identification framework, such as dividing the health status of self-propelled artillery into five levels of health, good, attention, deterioration and failure.

2) Select the key parameters reflecting the health status of self-propelled artillery

The selection of health parameters of self-propelled artillery is shown in the previous chapter.

3) Normalization processing of data

Health assessment of the self-propelled guns, the most important job is to process the data by normalization, assume that select n independent and effective reaction self-propelled guns health test parameters, in order to make these parameters can better describe the self-propelled guns health status, to the test values, respectively, and the fault test values, history and the last time test, compare the mean and standard values, therefore, the test data of normalized processing includes three items: this test data with the last time the breakdown test data comparison values, Comparison values of the test data with historical non-failure test mean and comparison values of the test data with standard data. Since the three comparison values are calculated in the same way, the normalization of the calculated test data and the historical non-fault test mean is illustrated below.

First of all, Calculated value deviation, the absolute value of the deviation between the current test data and the historical test data mean is calculated, If the test value of a parameter is X and the average value of the historical non-fault tests is X_L , then the deviation between the two is:

$$\delta_L = |x - x_L| \tag{7}$$

Secondly, select the normalized quantization function. According to the relationship between the test data and the mean value of the historical non-fault test, the semi-trapezoidal normalized quantization function was selected in this paper to calculate the health parameters more accurately, as shown in Figure 4.

Finally, calculate the normalized value. According to the normalized quantized trapezoidal function selected in the figure, the normalized value of the deviation between the test data and the mean value of the historical non-fault test data is

$$\lambda_L = \begin{cases} 1(\delta_L \le 0.3\delta_0) \\ \dfrac{\delta_0 - \delta_L}{0.7\delta_0}(0.3\delta_0 \succ \delta_L \prec \delta_0) \\ 0(\delta_L \succ \delta_0) \end{cases} \tag{8}$$

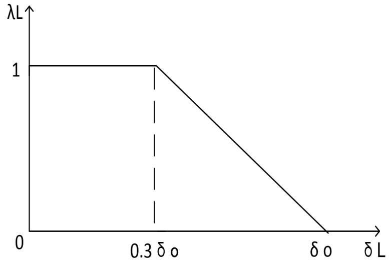


Figure 4. Half trapezoid normalized quantization function

Where, δ_0 is the max error limit. Using the same method, the normalized value λ_S of the test data compared to the last non-failure test data and the normalized value λ_B of the test data compared to the standard data can be obtained.

If $\lambda_L, \lambda_S, \lambda_B$, are all equal to 1, the self-propelled gun is "healthy," and don't exist Health hazard; If all the three are between 0.7 and 1, it means that the health status of self-propelled artillery is acceptable, and the average value of the health status index is taken. If any of the three is less than 0.7 and greater than 0, it indicates that there may be health risks. The health status index is the minimum value of three. If any of the three is 0, it indicates that self-propelled artillery is in a state of "disease". Therefore, the health status value of self-propelled artillery is:

$$\lambda = \begin{cases} 1(\lambda_L = \lambda_S = \lambda_B = 1) \\ \dfrac{\lambda_L + \lambda_S + \lambda_B}{3}(0.7 \le \lambda_L, \lambda_S, \lambda_B \prec 1) \\ \min(\lambda_L, \lambda_S, \lambda_B)(0 \prec \lambda_L or \lambda_S or \lambda_B \prec 0.7) \\ 0(\lambda_L or \lambda_S or \lambda_B = 0) \end{cases} \quad (9)$$

4) Determine the membership degree of parameter health status grade

According to the classification of the health status of the equipment, the health status, good status, attention status, deterioration status and failure status of the equipment are fuzzy, that is, due to the lack of an obvious transition from one health status level to another, the uncertainty is non-random and can be expressed by fuzzy set theory. The idea of fuzzy set is to fuzzy the absolute membership relationship in the classical set, so that the membership degree of an element to the set is no longer limited to 0 or 1, but can take any value on the interval $[0,1]$, which reflects the membership degree of an element to the set.

According to the test data, if the data is exceeded threshold value, the self-propelled gun can be judged to be ineffective. There are only fuzzy transition areas between state levels, with no clear boundaries. For example, self-propelled artillery on the edge of health-good state may be in both a healthy state and a good state, but the membership degrees of self-propelled artillery under the two states are different, so it is necessary to make a unified decision on the self-propelled artillery's health state and determine its health level. Since the normalized value of the test data is a representation of the health status of the parameters, the membership function of the parameters can be determined according to the normalized value of the test data. At the same time, due to the simple shape of the triangular membership function and the small difference from other more complex membership functions, this paper adopts the triangular membership function. According to the actual situation of equipment health degradation and expert experience, the triangular membership function of equipment parameters can be obtained, as shown in Figure 5.

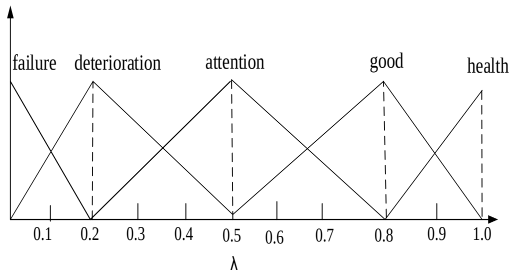


Figure 5. Membership function of fuzzy trigonometric functions

Can be seen from the figure 5, based on triangular membership functions, each param is affiliated with the health status of two adjacent levels, namely the health status of self-propelled guns parameter may belong to two adjacent healthy level in any one, but its membership may be different, and equipment belong to the adjacent two health level of the sum of membership degree of 1.

5) Calculate the weight of parameters

The weight is a measure to characterize the importance degree of evaluation index. To accurately evaluate the health status of self-propelled gun, the weight of each parameter should be determined. Since the normalized value of the test data represents the health state of the parameters, the smaller the normalized value is, the greater the deviation of the parameters from the standard value will be, and the worse their health state will be. Therefore, when evaluating the health status of the equipment, a few parameters with poor health status should be highlighted, that is, the worse the health status of the parameters, the smaller the normalized value and the greater the weight. In order to determine the weight according to the health state of the parameter, the objective weight of each parameter can be obtained by taking the reciprocal of the normalized value of each parameter and dividing the obtained result by the reciprocal of the normalized value of all parameters. The self-propelled gun has n parameters, the normalized value of $i(i=1,2......n)$ is λ_i , the weight of the parameter can be expressed as:

$$\omega_{i*}=\frac{\frac{1}{\lambda_i}}{\sum_{i=1}^{n}\frac{1}{\lambda_i}} \tag{10}$$

It can be seen from the formula that the smaller normalized value of the parameter is, the greater its weight will be. When the normalized value of a certain parameter is 0, it indicates that the test result of this parameter reaches the specified threshold value. At this time, the weight of this parameter is 1, while the weight of other parameters is 0. The health state of self-propelled gun can be judged directly according to the health state of this parameter, which is consistent with the actual situation.

6) *Calculate the basic trust allocation function*

When the D-S combination rule is applied to synthesize the health state of multiple parameters of the equipment, the synthesis formula of evidence theory considers that the importance of the evidence provided by all parameters is the same in the synthesis process. In fact, as one of the two parameters of the health status of serious deterioration, equipment integrated health status also fell sharply, namely the health status of equipment by a small number of the influence of the parameters of the poor state of health is larger, the evidence of each parameter in the process of evidence synthesis important degree is different, so it is necessary to introduce in the process of evidence synthesis can describe important evidence. The weight coefficient of degree is as follows:

After normalized by the weight formula in the above formula, the relative weight of the it's evidence parameter can be obtained as follows:

$$\omega_i=\frac{\omega_{i*}}{\max_{i=1...n}\omega_{i*}} \tag{11}$$

Set the maximum value of the relative weight parameter of evidence in this paper as 0.9, then the basic trust allocation function after weighted adjustment is:

$$\begin{cases} m_i(A_k)=\omega_i m_{i*}(A_k), i=1...n \\ m_i(\theta)=1-\sum_{k=1}^{N}m_i(A_k), k=1...n \end{cases} \tag{12}$$

Where, $m_i(A_k)$ is the basic trust allocation function before the weighted correction, $m_i(A_k)$ is the basic trust allocation function after the weighted correction, A_k is the single element focal element in the recognition framework, and N is the number of elements in the recognition framework.

7) *Determine the level of health through evidence fusion*

D-S (Dempster-Shafer) Evidence theory that uses the combination of Dempster's rule to integrate the knowledge or data of different experts or data sources, so that different descriptions of the same problem can be focused and one of them can be judged generative information has been widely used in information fusion, expert system, multi-attribute decision making and other fields.

According to the synthesis rules of D-S evidence theory, the basic trust allocation function m(A) of multiple evidence fusion is calculated by the following formula.

$$m(A) = \frac{1}{1-k} \sum_{\cap A_i = A} \prod_{1 \le i \le N} m_i(A_i) \qquad (13)$$

In the formula, $K = \sum_{\cap A_i = \Phi} m_i(A_i)$.

The state of health of self-propelled artillery is determined after the combination of evidence rules that is applied to the fusion of multiple evidence. After the evidence source correction is completed, information fusion can be carried out through the evidence theory, so as to obtain the health level of self-propelled artillery system. The specific process will be analyzed in the next chapter with examples.

IV. THE APPLICATION CASE

For example, as self-propelled guns, according to the test data, select five key parameters as test data for self-propelled guns' health status evaluation index, self-propelled guns has six test known, did not experience any maintenance, the sixth five parameters test results are qualified, in order to determine the health status of self-propelled guns, to assess the health status of the 6th test. First of all, according to the type (7), (8), (9), the test data of 5 test qualified parameters were normalized, can get the normalized value is (0.7817, 1.0000, 0.8087, 0.8534, 0.7688).

Since the normalized value of the test data is the representation of the health status of the parameters, in order to visually represent the health status of the five parameters of the equipment, a unit circle can be made and divided into 5 equal parts to obtain 5 radii. The length of each radius is 1, which is the maximum normalized value of each parameter test data, so that the normalized value of each parameter test data I can be represented as a point on the I radius, and the closer lambda I is to the center of the circle, the smaller the normalized value of the test data, the poorer the health of the parameters. By connecting the normalized values of the five parameters on the unit circle, the multi-parameter health status curve of the equipment can be obtained. Because its shape is similar to radar, it can be called health state radar chart, as shown in Figure 6.

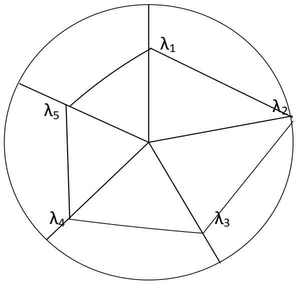


Figure 6. Health condition radar chart of parameters

The weight of each parameter should be determined according to Equation (10). The results are shown in Table 2.

TABLE II HEALTH STATUS ASSESSMENT RESULTS OF SELF-PROPELLED ARTILLERY PARAMETERS

Param	Normalized Value	Health Status Grade Membership					Weight
		Health	*Good*	*Attention*	*Deterioration*	*Failure*	
1	0.7817	0	0.939	0.061	0	0	0.214
2	1.0000	1	0	0	0	0	0.167
3	0.8087	0.0435	0.9565	0	0	0	0.206
4	0.8534	0.267	0.733	0	0	0	0.196
5	0.7688	0	0.896	0.104	0	0	0.217

In order to better represent the health status of the parameters, In order to analyze potential failures, after determining the membership degree of the parameter, that is, after assigning its basic probability, it can be known that the weight of the last one is the largest. Therefore, the weight of other parameters can be divided by the weight of the last parameter to determine its "discount rate". And according to Equations (11) and (12), the basic probability assignment after parameter modification is determined, as shown in Table 3

TABLE III BASIC PROBABILITY ASSIGNMENT AFTER PARAMETER MODIFICATION

Modified Parameters	**Health Status Grade Membership**					**Weight**
	Health	***Good***	***Attention***	***Deterioration***	***Failure***	
1	0	0.939	0.061	0	0	0.214
2	1	0	0	0	0	0.167
3	0.0435	0.9565	0	0	0	0.206
4	0.267	0.733	0	0	0	0.196
5	0	0.896	0.104	0	0	0.217

The basic probability assignment in Table 3 is synthesized by the evidence synthesis rule, and final synthesis result is $M = (0.0034, 0.9940, 0.0014, 0, 0, 0.0012)$ and. according to the principle of maximum membership degree, it can be known that the health state of self-guided artillery is "good".

In the same way, to assess the health status of self-propelled artillery from 2015 to 2020, it is known that the assessment results were healthy, healthy, healthy, good, and good. Considering that the state of the self-propelled gun is gradually degraded, the method proposed in this paper is feasible.

V. CONCLUSION

To evaluate the health status of self-propelled artillery, many indicators should be considered. In this paper, from the performance of the self-propelled guns, run indicators and environmental indicators from three aspects, the comprehensive and effective deputies elected to self-propelled guns health status of each index, and the state of the self-propelled guns can be divided into health, well being, attention, degradation state and failure state, better describe the health status of self-propelled guns. Secondly, the state of the parameters is evaluated. According to the measured data, the normalized value of each parameter is calculated, and the membership degree and weight of each parameter health state are determined. Finally, this paper adopts the fusion method of improved evidence theory and uses D-S evidence theory to fuse the data of various test parameters of health evaluation. The evaluation model of "either/or" is improved effectively, and the rationality of the evaluation model is verified through an example analysis, which is of great reference significance to the improvement of health evaluation methods for self-propelled artillery.

The method proposed in this paper can provide decision basis for health state assessment of self-propelled artillery and has certain reference significance for similar comprehensive health assessment. However, this paper only evaluates the health of the performance indicators, and it is advisable to consider the environmental indicators and operational indicators, as well as the entropy weight method to reduce the subjectivity of weight assignment, which needs to be further improved.

ACKNOWLEDGMENT

This research is supported by the Basic Natural Science Research Program of Shanxi Province (2019JM-603).

REFERENCES

[1] Weng Yan. State assessment of concrete cable-stayed Bridges based on analytic Hierarchy Process [D]. Chengdu: Southwest Jiao tong University,2017

[2] Zhang Yongqing et al. Evaluation of bridge safety by ANALYTIC Hierarchy Process [J]. Journal of Xi 'an Highway And Jiaotong University, 2012, 21(3):52-56.

[3] Hani G Melhem; Senaka Aturaliya. Bridge Condition Rating Using an Einenvector of Priority Settings [J]. Microcomputers in Civil Engineering, 2017, 11(3): 321-359.

[4] Li Wei. Research on status Assessment Method of HV Circuit Breakers Based on Fuzzy Comprehensive Evaluation [D]. Chongqing: Chongqing University, 2016.

[5] Xu Jiayun, He Xiaoming, ZHANG Jun. Application of Fuzzy Theory in bridge Evaluation [J]. Journal of Wuhan University of Technology, 2008, 25(7): 38-41

[6] Yager R R. Comparing approximate reasoning and probabilistic reasoning using the Dempster-Shafer framework [J]. International Journal of Approximate Reasoning, 2019, 50(5):8I2-821

[7] Basir O, Yuan X H. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory [J]. Information Fusion, 2017, 8(4):3 79-386.

[8] Lin T C. Partition belief median filter based on Dempster-Shafer theory for image processing [J]. Pattern Recognition, 2018, 41(1):139-151

[9] Wu W Z. Attribute reduction based on evidence theory in incomplete decision systems [J]. Infromation Sciences, 2018, 178(5):1355 一 1371

[10] Yin Ming, ye xiaohui, li qiaomin, wang xiaodong. Health assessment of radar transmitter based on evidence fusion [J]. Journal of naval engineering university, 2015, 27(05):48-51.

Adaptively Truncating Gradient for Image Quality Assessment

Minjuan Gao
School of Electrical and Control Engineering
Shaanxi University of Science & Technology
Xi'an, China
E-mail: gaominjuan1984@163.com

Xuande Zhang
School of Electronic Information and Artifificial Intelligence
Shaanxi University of Science & Technology
Xi'an, China
E-mail: 330374142@qq.com

Hongshe Dang
School of Electrical and Control Engineering
Shaanxi University of Science & Technology
Xi'an, China
E-mail: 783940896@qq.com

Abstract—**Objective image quality assessment (IQA) aims to develop computational models to predict the perceptual image quality consistent with subjective evaluations. As image information is presented by the change in intensity values in the spatial domain, the gradient, as a basic tool for measuring the change, is widely used in IQA models. However, does the change measured by the gradient actually correspond to the change perceived by the human visual system (HVS)? To explore this issue, in this paper, we analyze how the ability of the HVS to perceive changes is affected by the upper threshold, and we propose an IQA index based on an adaptively truncating gradient. Specifically, the upper threshold at each pixel in an image is adaptively determined according to the image content, and the adaptively truncating gradient is obtained by retaining the part of the gradient magnitude that is less than the upper threshold and truncating the part that is greater than the upper threshold. Then, the distorted image quality is calculated by comparing the similarity of the adaptively truncating gradient between a reference image and the distorted image. Experimental results on six benchmark databases demonstrate that the proposed index correlates well with human evaluations.**

Keywords-Image Quality Assessment; Human Visual System; Upper Threshold; Truncating Gradient

I. INTRODUCTION

Image quality assessment deals with the quantitative evaluation of the quality of images and can be widely used in image acquisition, compression, storage, transmission and other image processing systems. Generally, human beings are the ultimate receivers of images. Subjective evaluation by humans is a reliable IQA method, but it is cumbersome and difficult to apply in real-world scenarios. An objective IQA method aims to design mathematical models to automatically measure the image quality in a way that is consistent with human evaluations. According to the availability of ground-truth images, objective IQA indices fall into three categories: full-reference (FR), reduced-reference (RR) and no-reference (NR) models [1]. In this paper, the discussion is focused on FR models.

At present, there are two popular techniques for constructing FR models: knowledge-based and learning-based techniques. The deep learning method learns the evaluation model in an end-to-end manner, and its "black-box" lacks explanation. Furthermore, this approach requires a large number of training samples, but the cost of obtaining high-quality and

DOI: 10.21307/ijanmc-2020-034

convincing samples is relatively high. Currently, the commonly used method for obtaining samples is still data augmentation. In this work, we emphasize the knowledge-based approach, which uses knowledge about the HVS to heuristically construct IQA models. Investigating these models reveals that the gradient feature is widely employed. In analyzing the relationship between the gradient feature and the IQA task, the gradient has at least the following two characteristics. 1. The information contained in natural images is presented by changes in intensity value or color in the spatial domain. In extreme cases, the constant image (smoothness) and the pure noise image (variation in all directions) cannot convey any information. Thus, the feature of measuring change is widely used in IQA, with the gradient as the basic tool for measuring change. 2. The judgment of the image quality level in IQA is different from the classic discrimination task. The features for discrimination tasks, such as face recognition and fingerprint recognition, should be robust to image distortion, while the features for IQA should be sensitive to image distortion. The gradient feature is sensitive to image distortion and image content but is weak in robustness.

Representative FR models using the gradient feature include the feature similarity index (FSIM) [2], gradient magnitude similarity deviation index (GMSD) [3], superpixel-based similarity index (SPSIM) [4] and directional anisotropic structure metric (DASM) [5]. In the FSIM and GMSD, the image gradient magnitude is employed as the fundamental feature. SPSIM is computed on the basis of three features: superpixel luminance, superpixel chrominance and pixel gradient. The DASM is obtained by incorporating the gradient magnitude, anisotropy and local directivity features. Objective IQA models are designed by simulating the behaviors of the HVS, which integrates perception, understanding and assessing functions, that is, humans evaluate the image quality in the HVS perception space. Therefore, the features for IQA should be the subjective quantity perceived by the HVS. The gradient is often directly used in IQA models as an effective feature to measure change; however, does the change measured by the gradient actually correspond to that perceived by the HVS? In fact, the change measured by the gradient belongs to the objective quantity (objective physical stimulus), while that perceived by the HVS belongs to the subjective quantity (subjective response). Thus, how can one map the objective quantity to the subjective quantity? This mapping function is nonlinear, and it is difficult to accurately describe its form. Empirically, the ability of the human perception system to sense changes has a certain upper threshold. When the objective change exceeds the upper threshold, the subjective change increases insignificantly in situations such as the human perception of changes in salt-solution saltiness, at an outside temperature, and in the weight of objects carried.

In this paper, we discuss the ability of the HVS to perceive changes affected by the upper threshold by employing the adaptively truncating gradient to measure the change perceived by the HVS. We propose an IQA index based on the adaptively truncating gradient. Specifically, the upper threshold at each pixel in the image is adaptively determined according to the image content, and the adaptively truncating gradient is obtained by retaining the part of the gradient magnitude that is less than the upper threshold and truncating the part that is greater than the upper threshold. Experimental results on public databases show that the proposed index correlates well with the subjective judgments.

II. An IQA index based on adaptively truncating gradient

A. Definition of Adaptively Truncating Gradient

The image information is presented by the change in the intensity values in the spatial domain, and this change may be destroyed by degradation of the image quality. The gradient feature can effectively measure the change and is widely used in IQA algorithms. The image gradient can be obtained by convolving the image with a gradient operator, such as Sobel, Roberts and Scharr and Prewitt. Usually, a different gradient operator for the IQA model may yield distinguished performance. This problem was discussed in [2,6], where the experiment results showed that the Scharr operator can obtain a slightly better performance than the others. Here, we adopt a 3×3 Scharr operator whose templates along the horizontal (H) and vertical (V) directions take the following form:

$$\mathbf{h}_H=\frac{1}{16}\begin{bmatrix}3 & 0 & -3\\ 10 & 0 & -10\\ 3 & 0 & -3\end{bmatrix},\quad \mathbf{h}_V=\frac{1}{16}\begin{bmatrix}3 & 10 & 3\\ 0 & 0 & 0\\ -3 & -10 & -3\end{bmatrix}$$

Denote $\mathbf{r}=\left[r_1,\cdots,r_i,\cdots,r_N\right]$ for a reference image and $\mathbf{d}=\left[d_1,\cdots,d_i,\cdots,d_N\right]$for a distorted image, where i is the pixel index, and N is the number of total pixels. The image gradients in the horizontal and vertical directions can be obtained by convolution of the image with $\mathbf{h}_H$ and $\mathbf{h}_V$, and the gradient magnitude is computed from their root mean square. The gradient magnitudes of $\mathbf{r}$ and $\mathbf{d}$ at each pixel i, denoted as $G(\mathbf{r},i)$ and $G(\mathbf{d},i)$, are calculated as

$$G(\mathbf{r},i)=\sqrt{(\mathbf{r}\otimes\mathbf{h}_H)^2(i)+(\mathbf{r}\otimes\mathbf{h}_V)^2(i)} \qquad (1)$$

$$G(\mathbf{d},i)=\sqrt{(\mathbf{d}\otimes\mathbf{h}_H)^2(i)+(\mathbf{d}\otimes\mathbf{h}_V)^2(i)} \qquad (2)$$

Where the symbol ⊗ denotes the convolution operation.

The image gradient only reflects the objective changes in images. Since human evaluation of image quality is carried out in the HVS perception space, the image features extracted for IQA models should reflect the subjective changes perceived by the HVS. We consider that the ability of HVS to perceive changes is subject to the upper threshold. When the objective change exceeds the upper threshold, the subjective change does not obviously increase. In this study, we define the adaptively truncating gradient to measure the subjective change sensed by the HVS.

Denote T as the upper threshold. We define a truncating function $trunc(\cdot)$. For any given variable x, it is retained when it is less than T and truncated when it is greater than T. The specific expression is

$$trunc(x)=\begin{cases}T\,, & \text{if } x\geq T\\ x\,, & \text{if } x<T\end{cases} \qquad (3)$$

The truncating gradients of $\mathbf{r}$ and $\mathbf{d}$ at each pixel i are denoted as $G_T(\mathbf{r},i)$ and $G_T(\mathbf{d},i)$, and the upper threshold at this point is denoted as $T(i)$. Using formula (3), the calculation of $G_T(\mathbf{r},i)$ is as follows:

$$G_T(\mathbf{r},i)=trunc(G(\mathbf{r},i))=\begin{cases}T(i), & \text{if } G(\mathbf{r},i)\geq T(i)\\ G(\mathbf{r},i), & \text{if } G(\mathbf{r},i)<T(i)\end{cases} \qquad (4)$$

In Eq. (4), if the value of $G(\mathbf{r},i)$ is greater than $T(i)$, then $G(\mathbf{r},i)$ will be truncated, and the truncating gradient $G_T(\mathbf{r},i)$ is set to $T(i)$. That is, the part of the gradient magnitude that is greater than the upper threshold is masked. Otherwise, $G(\mathbf{r},i)$ is not be masked, and the truncating gradient $G_T(\mathbf{r},i)$ is set equal to $G(\mathbf{r},i)$. That is, the part of the gradient magnitude that is less than the upper threshold can be perceived by the HVS.

Similarly, using formula (3), $G_T(\mathbf{d},i)$ is calculated as follows:

$$G_T(\mathbf{d},i)=trunc(G(\mathbf{d},i))=\begin{cases}T(i), & \text{if } G(\mathbf{d},i)\geq T(i)\\ G(\mathbf{d},i), & \text{if } G(\mathbf{d},i)<T(i)\end{cases} \qquad (5)$$

Obviously, for the calculation of the truncating gradients $G_T(\mathbf{r},i)$ and $G_T(\mathbf{d},i)$ in Eq. (4) and (5), the selection of the upper threshold $T(i)$is very important. According to Weber’s law, the ratio of the stimulus change that causes a just noticeable difference (JND) from the original stimulus intensity is a constant. In psychology, the HVS has the property of light adaptation, and the perception of luminance obeys Weber’s law [7]. The just noticeable incremental luminance over the background by the HVS is related to the background luminance.

Inspired by this recognition, in contrast to Weber’s law, we consider that the upper threshold for truncating the significantly perceptible stimulus change is also related to the original stimulus intensity value. Because different pixels in the image correspond to different gray values, the original stimulus intensity values will also be different. Here, we adaptively determine the upper threshold according to the background luminance of different areas of the image.

The adaptively upper threshold is defined as

$$T(i)=\frac{I(i)}{T_0} \qquad (6)$$

Where T_0 is an adjustable threshold parameter. (The details of selecting T_0 will be presented in section III-A.) $I(i)$ takes the larger value of the luminance of **r** and **d** at point *i*.

$$I(i)=max(\bar{r}(i),\bar{d}(i)) \qquad (7)$$

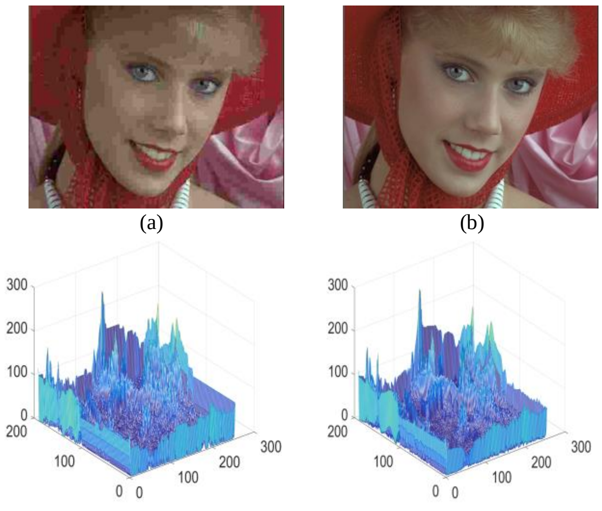
In formula (7), the luminance values $\bar{r}(i)$ and $\bar{d}(i)$ at pixel i of $\mathbf{r}$ and $\mathbf{d}$ is estimated by formulas (8) and (9). For reference image $\mathbf{r}$, denote the square neighborhood as $\boldsymbol{\Omega}_i^r$ with center of pixel i and radius of t, and let the intensity value of any pixel in the neighborhood be $r_{i,j}$, $j \in \mathbf{\Omega}_i^r$. Similarly, for the distorted image, denote the square neighborhood as $\mathbf{\Omega}_i^d$ with center of pixel i and radius of t, and let the intensity value of any pixel in the neighborhood be $d_{i,j}$, $j \in \mathbf{\Omega}_i^d$.

$$\bar{r}(i)=\frac{1}{m}\sum_{j=1}^{m} r_{i,j} \tag{8}$$

$$\bar{d}(i)=\frac{1}{m}\sum_{j=1}^{m} d_{i,j} \tag{9}$$

Where $m=(2t+1)^2$.

Based on Eq. (6), the value of the upper threshold at each pixel in an image can be adaptively determined according to the image content. Then, the adaptively truncating gradient is obtained by formulas (4) and (5). Figure 1 shows the gradient map and the adaptively truncating gradient map corresponding to the reference image and the distorted image. It can be seen that the maximum amplitude of the gradient map is approximately 250, while the maximum amplitude of the adaptively truncating gradient is approximately 70.

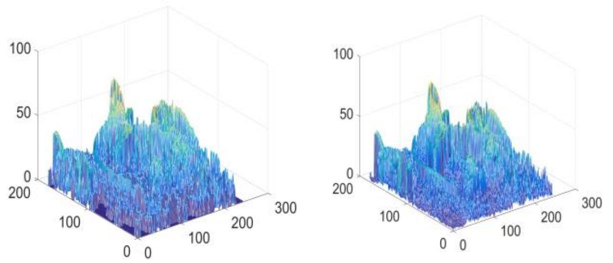


(a)

(b)

(c)

(d)



(e)

(f)

Figure 1. The gradient map and the adaptively truncating gradient map corresponding to the reference image and the distorted image. (a) the reference image. (b) the distorted image. (c) and (d) are the gradient map of (a) and (b) , respectively. (e) and (f) are the adaptively truncating gradient map of (a) and (b) , respectively.

B. *The Proposed IQA Index*

With the adaptively truncating gradient defined, the local quality of the distorted image is predicted by the similarity between the adaptively truncating gradient of **r** and **d**, which is defined as

$$S(i)=\frac{2G_T(\mathbf{r},i)\cdot G_T(\mathbf{d},i)+C}{G_T^2(\mathbf{r},i)+G_T^2(\mathbf{d},i)+C} \tag{10}$$

Where the parameter C is introduced to avoid the denominator becoming zero and supplies numerical stability. The range of $S(i)$ is from 0 to 1. Obviously, on the one hand, $S(i)$ is close to 0 when $G_T(\mathbf{r},i)$ and $G_T(\mathbf{d},i)$ are quite different. On the other hand, $S(i)$ will achieve the maximal value 1 when $G_T(\mathbf{r},i)$ is equal to $G_T(\mathbf{d},i)$.

The overall quality score of the distorted image is predicted by the local quality $S(i)$, which is calculated as follows :

$$score=\frac{1}{N}\sum_{i=1}^{N} S(i) \tag{11}$$

A higher score indicates better image quality.

III. EXPERIMENTAL RESULTS

A. *Experimental Setup*

All the experiments in this study were implemented in MATLAB R2016b and executed on a Lenovo Ideapad700 laptop with Intel Core i5-6300HQ@2.3-

GHz CPU and 4 GB RAM. Several well-known FR metrics were used when comparing performances with the proposed method, including PSNR, SSIM[1], FSIM [2], GMSD[3], DASM[5], IFC [8], VIF [9], MS-SSIM [10], and SSRM [11]. To widely evaluate the performance of these metrics, six public databases were employed for the experiments: TID2013 [12], TID2008 [13], CSIQ [14], LIVE [15], IVC [16] and A57 [17]. The TID2008 database consists of 25 reference images and a total of 1700 distorted images, each of which is distorted using 17 different types of distortions at four different levels of distortion. The TID2013 is an expanded version of TID2008, which contains 3000 distorted images with 24 distortion types. The LIVE database includes 29 reference images and 779 distorted images with five distortion types. The CSIQ database contains 30 original images and 886 distorted images degraded by six types of distortion. The IVC database consists of 10 reference images and 185 distorted images. The A57 database includes 3 reference images and 54 distorted images. Note that for the color images in these databases, only the luminance component is evaluated.

Four commonly used performance criteria are employed to evaluate the competing IQA metrics. The Spearman rank order correlation coefficient (SROCC) and Kendall rank order correlation coefficient (KROCC) are adopted for measuring the prediction monotonicity of an objective IQA metric. For compute the other two criteria, the Pearson linear correlation coefficient (PLCC) and the root mean squared error (RMSE), we need to apply a regression analysis. The PLCC measures the consistency between the objective scores after nonlinear regression and the subjective mean opinion scores (MOS). The RMSE measures the relative distance between the objective scores after nonlinear regression and MOS. For the nonlinear regression, we used the following mapping function:

$$Q_P = \beta_1\left[\frac{1}{2} - \frac{1}{1+e^{\beta_2(Q-\beta_3)}}\right] + \beta_4 Q + \beta_5 \qquad (12)$$

where Q and Q_P are original objective scores of an IQA metric and the objective scores after regression, respectively. β_i, $i = 1, 2, \cdots, 5$ are the fixed parameters. Higher values of SROCC, KROCC, PLCC and lower RMSE values indicate a better performance of IQA metrics.

For the proposed metric, there are three parameters that need to be set to obtain the final quality score. They are T_0, t and C. Selecting the first 8 reference images and corresponding 544 distorted images in the TID2008 database as the testing subset, we choose the parameters that can yield the highest SROCC. The result is $T_0 = 3$, $t = 51$ and $C = 1600$.

To further analyze the effect of threshold parameter T_0, more experiments were carried out. Figure 2 shows the SROCC performance with different T_0 values on six databases. On most databases, SROCC can is best when T_0 is 3. This result indicates that the range of upper threshold T is approximately [0,255/3] for an 8-bit grayscale image according to formula (6). If the change in image intensity is above 255/3, then it will be masked in visual perception.

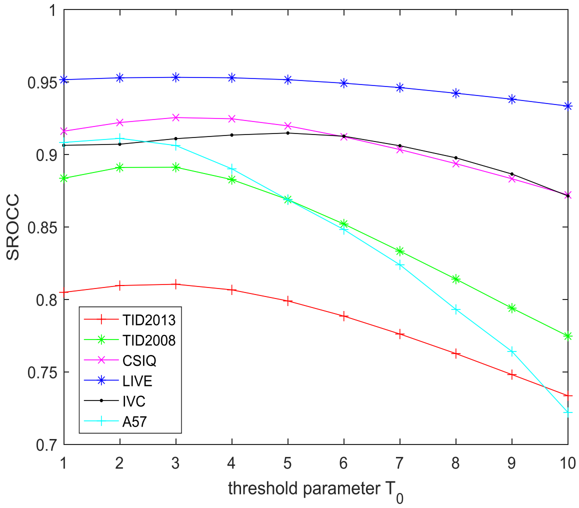


Figure 2. SROCC performance with different T_0 values on six databases

B. *Performance Comparison*

Table Ⅰ lists the SROCC, KROCC, PLCC and RMSE results of ten metrics on six databases, and the two best results of each row are highlighted in bold. Overall, the methods which employed the gradient feature performs well across all the databases, such as FSIM, GMSD, DASM and the proposed metric. This partly demonstrates the validity of considering the degradation of gray changes in quality evaluation. Furthermore, the proposed metric performs well, outperforming SSIM and SSRM and competing with FSIM and GMSD.

TABLE I. COMPARISON THE PERFORMANCE RESULTS OF TEN IQA METRICS ON SIX PUBLIC DATABASES. THE FIRST TWO ARE MARKED IN BOLD

Database	**criteria**	**PSNR**	**SSIM (2004)**	**MS-SSIM (2003)**	**IFC (2005)**	**VIF (2006)**	**FSIM (2011)**	**GMSD (2014)**	**DASM (2017)**	**SSRM (2018)**	**Proposed**
TID2013	SROCC	0.6396	0.7417	0.7859	0.5389	0.6769	0.8015	**0.8038**	0.8025	0.7500	0.8105
	KROCC	0.4698	0.5588	0.6047	0.3939	0.5147	0.6289	**0.6334**	0.6321	0.5718	0.6387
	PLCC	0.7017	0.7895	0.8329	0.5538	0.7720	**0.8589**	0.8542	0.8574	0.8078	0.8601
	RMSE	0.8832	0.7608	0.6861	1.0322	0.7880	**0.6349**	0.6444	0.6547	0.7307	0.6324
TID2008	SROCC	0.5531	0.7749	0.8542	0.5675	0.7491	0.8805	**0.8906**	-	0.8332	0.8913
	KROCC	0.4027	0.5768	0.6568	0.4236	0.5860	0.6946	**0.7090**	-	0.6535	0.7042
	PLCC	0.5734	0.7732	0.8451	0.7340	0.8090	0.8738	**0.8786**	-	0.8379	0.8745
	RMSE	1.0994	0.8511	0.7173	0.9113	0.7888	0.6525	**0.6408**	-	0.7324	0.6458
LIVE	SROCC	0.8756	0.9479	0.9513	0.9259	**0.9636**	**0.9634**	0.9546	0.9601	0.9608	0.9531
	KROCC	0.6865	0.7963	0.8045	0.7579	0.8282	**0.8337**	0.8237	0.8218	**0.8312**	0.8211
	PLCC	0.8721	0.9449	0.9430	0.9248	**0.9598**	0.9597	0.9515	0.9571	**0.9695**	0.9379
	RMSE	13.368	8.9455	9.0956	10.392	7.6734	7.6780	**7.1131**	7.7716	**5.6639**	8.0188
CSIQ	SROCC	0.8058	0.8756	0.9133	0.7671	0.9195	0.9242	**0.9571**	**0.9523**	0.9369	0.9251
	KROCC	0.6084	0.6907	0.7393	0.5897	0.7537	0.7567	**0.8122**	**0.8041**	0.7791	0.7575
	PLCC	0.8001	0.8613	0.8998	0.8381	0.9277	0.9120	**0.9543**	**0.9531**	0.9097	0.9055
	RMSE	0.1575	0.1334	0.1145	0.1432	0.0980	0.1077	**0.0791**	**0.0799**	0.1138	0.1114
IVC	SROCC	0.6884	0.9018	0.8980	0.8993	0.8964	**0.9262**	0.8789	0.8966	0.9047	**0.9103**
	KROCC	0.5218	0.7223	0.7203	0.7202	0.7158	**0.7564**	0.6882	0.7179	0.7310	**0.7352**
	PLCC	0.7199	0.9119	0.8934	0.9080	0.9028	**0.9376**	0.8549	0.9190	0.9132	**0.9206**
	RMSE	0.8456	0.4999	0.5474	0.5105	0.5239	**0.4236**	0.6320	0.5220	0.4965	**0.4758**
A57	SROCC	0.6189	0.8066	0.8394	0.3185	0.6223	**0.9181**	0.9103	**0.9215**	0.8527	0.9062
	KROCC	0.4309	0.6058	0.6478	0.2378	0.4589	**0.7639**	0.7513	**0.7782**	0.6604	0.7289
	PLCC	0.6587	0.8017	0.8504	0.4548	0.6158	**0.9252**	0.9085	**0.9429**	0.8528	0.8530
	RMSE	0.1849	0.1469	0.1293	0.2189	0.1936	0.0933	01027	**0.0813**	**0.0707**	0.1283

TABLE II. COMPARISON SROCC FOR INDIVIDUAL DISTORTION OF TEN IQA METRICS ON TID2013 DATABASE. THE FIRST TWO ARE MARKED IN BOLD

Database	**Distortion type**	**PSNR**	**SSIM (2004)**	**MS-SSIM (2003)**	**IFC (2005)**	**VIF (2006)**	**FSIM (2011)**	**GMSD (2014)**	**DASM (2017)**	**SSRM (2018)**	**Proposed**
TID2013	Awgn	0.9291	0.8671	0.8646	0.6612	0.8994	0.8973	**0.9461**	**0.9299**	0.8545	0.9293
	Awgn-color	**0.8986**	0.7726	0.7730	0.5352	0.8299	0.8208	**0.8689**	08612	0.7757	0.8463
	Spatial-correlated	0.9197	0.8515	0.8544	0.6601	0.8835	0.8750	**0.9348**	**0.9301**	0.8392	0.9178
	Mask-noise	**0.8321**	0.7767	0.8073	0.6932	**0.8450**	0.7944	0.7085	0.8019	0.8184	0.8068
	HF-noise	0.9140	0.8634	0.8604	0.7406	0.8972	0.8984	**0.9164**	**0.9179**	0.8754	0.9069
	Impulse-noise	**0.8969**	0.7503	0.7629	0.6408	0.8537	0.8072	0.7633	**0.8550**	0.7872	0.8336
	Quantization-noise	0.8801	0.8657	0.8706	0.6282	0.7854	0.8719	**0.9057**	**0.9032**	0.8496	0.8629
	GB	0.9155	0.9668	0.9673	0.8907	0.9650	0.9551	0.9114	0.9546	**0.9674**	**0.9686**
	Denoising	0.9481	0.9254	0.9268	0.7779	0.8911	0.9302	**0.9525**	**0.9496**	0.9288	0.9361
	JPEG	0.9189	0.9200	0.9265	0.8357	0.9192	0.9324	**0.9500**	0.9473	0.9287	**0.9515**
	JP2K	0.8840	0.9468	0.9504	0.9078	0.9516	0.9577	**0.9656**	0.9620	0.9562	**0.9635**
	JPEG-trans-error	0.7682	0.8493	0.8475	0.7425	0.8409	0.8464	0.8401	**0.8534**	0.8369	**0.8802**
	JP2K-trans-error	0.8886	0.8828	0.8889	0.7769	0.8761	0.8913	**0.9135**	0.8966	0.8765	**0.9141**
	Pattern-noise	0.6864	0.7821	0.7968	0.5737	0.7720	0.7917	**0.8143**	0.8138	0.7745	0.7632
	Block-distortion	0.1552	0.5720	0.4801	0.2414	0.5306	0.5489	**0.6630**	0.6338	0.3186	**0.6635**
	Mean-shift	0.7671	**0.7752**	**0.7906**	0.5522	0.6276	0.7531	0.7356	0.6127	0.6919	0.6143
	Contrast change	0.4416	0.3775	0.4634	-0.1798	**0.8386**	0.4686	0.3253	0.3498	0.4519	**0.4889**
	Saturation change	**0.0944**	-0.4141	-0.4099	-0.4029	-0.3099	-0.2748	-0.1907	**0.0382**	-0.2513	-0.2602
	Multiple-noise	**0.8911**	0.7803	0.7786	0.61423	0.8468	0.8469	**0.8880**	0.8814	0.8067	0.8698
	Comfort-noise	0.8410	0.8566	0.8528	0.81620	0.8946	0.9121	**0.9298**	**0.9203**	0.8921	0.9112
	Noisy-compression	0.9144	0.9057	0.9068	0.8180	0.9204	**0.9466**	**0.9631**	0.9402	0.9164	0.9367
	Color quantization	**0.9269**	0.8542	0.8555	0.6006	0.8414	0.8760	0.9098	**0.9177**	0.8546	0.8952
	Chromatic abbr.	**0.8871**	0.8775	0.8784	0.8210	0.8848	0.8715	0.8517	0.8693	0.8844	**0.8849**
	Sparse sample	0.9044	0.9461	0.9483	0.8885	0.9353	0.9565	**0.9684**	**0.9669**	0.9541	0.9601

Among the six databases, TID2013 has the highest number of distorted types. Table Ⅱ lists the SROCC results of ten metrics about each individual distorted type of the TID2013 database. The proposed algorithm performs well in variety of distortion types. In particular, the proposed algorithm is outstanding for JPEG, JP2K and JPEG-trans-error distortion types that are sensitive to variations.

IV. CONCLUSION

In this paper, we discuss the problem of whether the change measured by the gradient correspond to the change perceived by the HVS. Considering that the ability of the HVS to perceive changes is affected by the upper threshold, we defined the adaptively truncating gradient and proposed a novel IQA index. Numerical experimental results showed that this index performs well on multiple databases. In addition, more studies need to be conducted to address this problem due to its complexity. In future research, we expect to using machine learning methods to further understand this issue.

ACKNOWLEDGMENT

This work was supported by the Natural Science Foundation of Shaanxi Province under Grant 2020JM-509.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[2] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[3] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly effificient perceptual image quality index,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, Feb. 2014.

[4] W. Sun, Q. Liao, J. Xue, and F. Zhou, “SPSIM: A superpixel-based similarity index for full-reference image quality assessment,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4232–4244, Sept. 2018.

[5] L. Ding, H. Huang, and Y. Zang, “Image quality assessment using directional anisotropy structure measurement,” IEEE Trans. Image Process., vol. 24, no. 4, pp. 1799–1809, Apr. 2017.

[6] X. Zhang, X. Feng, W. Wang, and W. Xue, “Edge strength similarity for image quality assessment,” IEEE Signal Process. Lett., vol. 20, no. 4, pp. 319–322, Apr. 2013.

[7] Z. Wang and A. C. Bovik, “Bottom-up approaches for full-reference image quality assessment ,” in Modern image quality assessment, Vermont, VT, USA: Morgan and Claypool, 2006, pp. 17–40.

[8] H. R. Sheikh, A. C. Bovik, and G. de Veciana, “An information fifidelity criterion for image quality assessment using natural scene statistics,” IEEE Trans. Image Process., vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[9] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., vol. 15, no. 12, pp. 430–444, Feb. 2006.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multi-scale structural similarity for image quality assessment,” in Proc. IEEE Asilomar Conf. Signals, Syst. Comput., Nov. 2003, pp. 1398–1402.

[11] A. Ahar, A. Barri and P. Schelkens, "From Sparse Coding Significance to Perceptual Quality: A New Approach for Image Quality Assessment," IEEE Trans. Image Process., vol. 27, no. 2, pp. 879-893, Feb. 2018.

[12] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J, Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo, “Color image database TID2013: Peculiarities and preliminary results,” in Proc. 4th Eur. Workshop Vis. Inf. Process., Jun. 2013, pp. 106–111.

[13] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, “TID2008A database for evaluation of full-reference visual quality assessment metrics,” Adv. Modern Radioelectron., vol. 10, pp. 30–45, May. 2009.

[14] C. Larson and D. M. Chandler, Categorical Image Quality (CSIQ) Database 2009 [Online]. Available: http://vision.okstate.edu/csiq

[15] H. R. Sheikh, K. Seshadrinathan, A. K. Moorthy, Z. Wang, A. C. Bovik, and L. K. Cormack, Image and Video Quality Assessment Research at LIVE 2004 [Online]. Available: http://live.ece.utexas.edu /research /quality

[16] A. Ninassi, P. Le Callet, and F. Autrusseau, Subjective Quality Assessment IVC Database 2005 [Online]. Available: http://www2.irccyn. ecnantes.fr/ivcdb

[17] D. M. Chandler and S. S. Hemami, A57 Database 2007 [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.htm

Discussion on Decimal Network Based on IPV9

Hu Shun
Guilin University of Electronic Technology

Xu Dongmei, Gao Lin
China Institute of Electronic Technology Standardization

Abstract—**This paper introduces the core technology of decimal network digital domain name and IPV9 protocol family, and analyzes the technical characteristics of decimal network. Three kinds of common network transition techniques are listed, and various problems of decimal network application are discussed.**

Keywords-Decimal Network; Digital Domain Names; IPV9

I. INTRODUCTION

TCP/IP network architecture and Protocol standards in recent years, computer network research and application of hot technology. At present, the widely used IP protocol is IPv4, based on which the Internet has become the largest computer network system in the world. However, with the rapid development of economic globalization and modern communication technology and network, the scale of computer network is expanding rapidly, and IPv4 protocol starts to expose various problems. Such as: IP address resources, address allocation efficiency is low, no consideration of confidential transmission. Facts show that IPv4 cannot meet the requirements of future Internet development. In this context, countries around the world have stepped up the work of the next generation of Internet protocols. IPv6 has been selected as an international standard by the Internet Engineering Task Force (IETF), while IPV9 proposed in 1992 was abandoned by the ETF due to its large address. Later, with the introduction of digital domain name system (DDNS), gradually developed into a 256-bit address IPV9 decimal network with China's independent intellectual property rights.

II. DIGITAL DOMAIN NAME TECHNOLOGY

The so-called digital domain name, refers to the Arabic numerals (0~9) as the Internet intelligent terminal domain name. The coding of digital domain name refers to the telephone coding rules. It adopts a hierarchical structure according to different regions and consists of root, country/region, and city and user code from top to bottom.

Digital domain names provide users with an alternative to English domain names. At the same time, they have the following characteristics:

- Use of class telephone Numbers to facilitate domain name management and division;
- Make it easier and faster to browse the Internet on smart terminals in the future;
- Provides conditions for the realization of network end-to-end communication in the future;
- Number resources can be integrated to facilitate the integration of the three networks.

Decimal network introduced the digital domain name system, and can be compatible with English domain name, Chinese domain name system. Through

DOI: 10.21307/ijanmc-2020-035

the DNS of the domain name server, the digital domain name entered by the user is converted into the corresponding IP address to achieve the purpose of accessing the host. Currently, DDNS maps digital domain names to dynamic IP addresses by installing a small program on the client side. When the user dial-up Internet access, the user will be dynamic IP address and user's digital domain name information notification server, the server will be the user's digital domain name and dynamic IP address registered in the DDNS resolution system, and then began to provide digital domain name resolution services. When the user is offline, the user's digital domain name information is removed from the DDNS resolution system.

At present, digital domain name system and IPV9 protocol has been recognized by some countries and regions. China has developed some IPV9 related network equipment and systems, solved the problem of interconnection between IPv4 network and IPV9 network, and realized the independent function of domain name resolution, domain name allocation, IP address allocation and MAC address allocation. After several years of trial operation, the experimental system established in Jinshan County, Changning District and Fujian province has been successfully tested in five small projects. The Shanghai experimental area is connected to IPv4 networks in Beijing and Hangzhou by tunnel. In addition, various applications based on THE IPV9 decimal network have been developed or are being developed.

III. IPV9 PROTOCOL FAMILY

IPV9 protocol family is a decimal network base protocol, including IPV9 header protocol, address protocol and transition protocol.

A. IPV9 header Protocol

IPV9 packet header format and field meaning are specified, including basic header and extended header.

1) Basic headers

The basic header format specified in the IPV9 header protocol is shown in Table 1.

TABLE I. IPV9 HEADER FORMAT

version	category	Flow label	Payload length	Under one head	Hop limit
Source address					
The destination address					

Version: The length is 4 bits, indicating the protocol version number.

Category: Length 8 bits, 0 to 15 as priority values. The priority classes 0 through 7 are used to specify communication Settings and are used by packet senders to control traffic. 8 to 15 is used to specify traffic that will not fall back in the event of congestion. 16 and 17 assign audio and video, called absolute values, to ensure uninterrupted transmission of audio and video. Others are reserved values.

- Stream label: A length of 20 bits, used to identify packets belonging to the same traffic.
- Net charge length: The length is 16 bits, indicating the number of bytes of the packet behind the IPV9 header.
- Next header: The length is 8 bits, which indicates the protocol type in the header field that follows the IPV9 header.

- Jump limit: The length is 8 bits, indicating the maximum number of times the packet can be forwarded by the node.
- Source address: The length is 256bit, representing the sender address.
- Destination address: The length is 256bit, representing the recipient address.

2) Extended headers

Between the packet IPV9 header and the high-level protocol header, there may be specialized headers, called extended headers, to represent optional Internet layer information. The number of extended headers is small, and each is identified by a different next header value. The IPV9 packet can come with no or multiple extended headers, which need to be defined by the next header field in the previous header. The extended header of IPV9 protocol includes six types: segment selection, route selection, segmentation, destination options, identification and encapsulation of security payloads.

According to the IPV9 header protocol, IPV9 header adopts the form of basic header + extended header chain. Compared to THE IPv4 header, the IPV9 header has removed the header length field and replaced the Type of Service field with the Traffic Class field. The Protocol Type and time-to-live (TTL) fields have been renamed and slightly modified. In addition, the Flow Label field has been added.

Although the total length of the IPV9 base header is nearly four times that of the default IPv4 header (20 bytes), to 72 bytes, it is actually simplified. Because the vast majority of the header is occupied by two 64-byte IPV9 addresses, the rest of the header takes up only eight bytes.

B. IPV9 Address Protocol

The IPV9 address protocol specifies that the IPV9 address is 256 bits, enabling a large addressing space of 2256. According to the data transmission mode, it can be divided into unicast, on-demand and multicast. In addition, the addressing model of IPV9, text representation of IPV9 address, text representation of address prefix, address type representation, monocular address, multiple destinations address and other contents are also specified.

- IPV9 addressing model: Specifies that all types of IPV9 addresses are assigned to interfaces, not nodes.
- Textual representation of 1PV9 addresses: Specifies that IPV9 addresses use "bracket decimal" notation. IPV9 addresses can be divided into pure IPV9 addresses, IPv4-compatible IPV9 addresses, ipV6-compatible IPV9 addresses, special compatible addresses, full decimal addresses, and transitional IPV9 addresses. For convenience of reading, some abbreviations are specified for text representation of addresses.
- Textual representation of address prefixes: Address prefixes for IPV9 addresses are specified to reflect the network hierarchy.
- Address type representation: Specifies some of the high boot bits of the IPV9 address as the format prefix FP to indicate different types of IPV9 addresses. The length of these format prefixes varies.
- Monocular address: Represents a single network interface. Messages addressed to monocular address will be sent to the unique network interface identified by it. The forms of monocular addresses specified in the IPV9 address protocol include the aggregated global monocular address, the decimal Internet address, the domain name decision and assignment organization address, the IPX address, the local IPV9 monocular address, and the IPv4 compatible address.

- Multiple destinations address: A class of IPV9 addresses assigned to multiple network interfaces at the same time. The IPV9 address protocol states that multiple destinations address addresses are assigned from single-mesh addresses, using the same format as single-mesh addresses. When a monocular address is assigned to multiple network interfaces, it is functionally converted to a multiple destinations address.

C. *IPV9 Interim Agreement*

The IPV9 transition protocol specifies the header format of the IPV9 transition and the definition of the address text representation, addressing model, and node address, including a detailed description of the current transition header and address format defined.

The transitional headers used the original IPv4 header, only changing the version number to 9 to distinguish it from the original IPv4 header. The transitional address adopts the latter two segments of the IPV9 address, a total of 64 bits.

IV. TECHNICAL FEATURES OF IPV9 DECIMAL NETWORK

A. *Address space*

The wealth of IP address resources is undoubtedly an important advantage of the IPV9 decimal network. Due to the 256-bit address, the theoretical address capacity is 2256, which is said to be able to assign a permanent address to the world's human needs for 750 years, and can be automatically increased sequentially after 750 years. So the address is large enough to solve the IPv4 address resource constraints.

B. *Digital domain name System*

Digital domain name is an important part of IPV9 decimal network system, which is compatible with English domain name and Chinese domain name. It is impossible to replace English domain names, but it is a good choice for users who are not used to English domain names. In addition, due to digital domain name technology, the decimal network system can be the domain name, IP address, MAC address unified representation into decimal text.

C. *Automatic configuration*

According to IPV9 plug and play data, IPV9 supports stateless and stately host address automatic configuration, which USES DHCP of IPV9.

D. *Security*

The special encryption mechanism is adopted to ensure the safe transmission of network data.

E. *Mobility support*

The IPV9 decimal network establishes an IPV9 tunnel between the mobile unit and the home agent, and then relays the packets to the mobile unit's home address received by the "proxy" of the mobile unit through the tunnel to the current location of the mobile unit, so as to realize the support for network terminal mobility.

IPV9 decimal network introduced the digital domain name technology, convenient digital button Internet, simplified the difficulty of network management. The expansion of address space and the introduction of security mechanisms have solved the problems faced by IPv4 networks. Support for QoS, automatic configuration, and mobility enables it to better meet multiple business requirements. IPV9 protocol can theoretically meet the requirements of the new generation of Internet development. At present, after the experimental verification stage, completed the development of basic hardware equipment, has entered the stage of small-scale application.

V. TRANSITION TECHNOLOGY FROM IPV4 TO IPV9

Although IPV9 has many technical features and can solve various problems faced by IPv4, IPV9 has a long history. The IPV9 decimal network is not likely to replace the huge IPv4 network in a short time, but will go through a long process of coexistence and transition. Drawing on a number of IPv6 technologies, lPv9 also supports the IETF Next Generation Internet Transition Working Group to propose dual stack, tunneling technology, and NAT-PT (Network address translation/protocol translation).

A. *Dual protocol stack technology*

This is the simplest way to handle transition problems. This mechanism enables the device to handle both types of protocols by running both 1PV9 and IPv4 stacks on a single device, as shown in Table 2.

TABLE II. STRUCTURE DIAGRAM OF DOUBLE PROTOCOL STACK

The application layer	
Transport Layer (TCP/UDP protocol)	
IPv4	IPV9
Network interface layer	

The dual stack mechanism is intuitive and easy to understand. The problem is that the dual stack still requires the corresponding host to configure IPv4 addresses, otherwise it is invalid, which goes against the original intention of using IPV9 to solve the problem of insufficient IPv4 addresses. In practice, it is impossible for all hosts or terminals to upgrade to support dual stacks, and using dual stacks will increase the burden on hosts and reduce performance. Therefore, the application must be combined with other technologies.

B. *Tunneling Technology*

Tunneling provides a way to pass IPV9 data over existing IPv4 routing systems, as shown in Figure 3. The technical principle is that at the entrance of the tunnel, the router encapsulates the 1PV9 data packet into the IPv4 packet, whose source address and destination address are the IPv4 addresses of the tunnel entrance and exit respectively. The encapsulated IPv4 packet will be transmitted through the IPv4 routing body, and the protocol domain of the packet header is set to 141. Indicates that the load of this packet is an IPV9 packet, which is taken out and forwarded to the destination station at the exit of the tunnel.

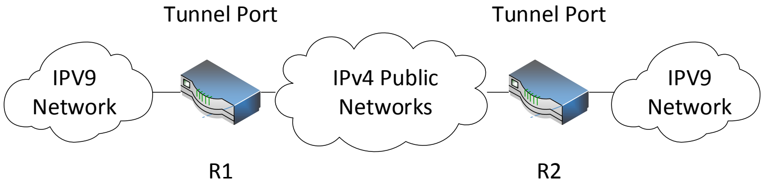


Figure 1. IPV9 over IPv4 tunnels

Tunneling technology requires modifications only at the entrance and exit of the tunnel, with no other requirements, and is therefore very easy to implement. It is currently the most widely used transition technology, the disadvantage is that IPV9 host and IPv4 host cannot achieve direct communication. Transitional IPV9 decimal network supports two tunnel technologies: IPV9 over IPv4 and IPv4 over IPV9, which can be divided into automatic configuration and manual configuration according to address configuration. The improved technology includes tunnel agent technology.

C. *NAT - PT*

Nat-PT technology is a protocol translation technology that performs both header and semantic translation (PT) between IPv4 and IPV9 packets while performing IPv4/IPV9 address translation (NAT). Through the introduction of Nat-PT router, the intercommunication between IPv4 sub-net and IPV9 sub-net can be realized. The network structure is shown in Figure 2.

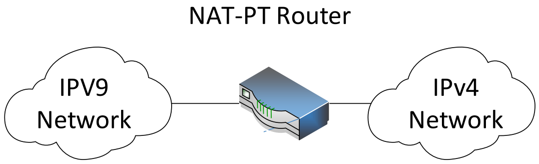


Figure 2. Nat-PT network structure diagram

Nat-PT can solve the problem of tunnel technology and realize direct communication between IPV9 and IPv4 host, which is suitable for the initial stage of the transition of IPV9 network with small scale.

VI. CONCLUSION

The decimal network is based on the digital domain name and IPV9, two core technologies independently developed in China, with independent intellectual property rights. At the same time, it can solve various problems faced by the existing IPv4 network and meet the requirements of the development of the next generation of Internet. If it can be applied and popularized, it will promote the great development of the Internet and help China get rid of the situation of being restrained by others in Internet technology. However, the establishment and promotion of standards is not a simple technical issue, which involves the interests of various countries and groups. Therefore, the popularization of decimal network cannot be accomplished overnight, so it should be supported by the state and applied in government departments, military departments and other departments with higher requirements on network performance and security. And then gradually spread to achieve business operations. At the same time, it is also necessary to coordinate the relationship between domestic and foreign interest groups and promote their international standardization process, so as to finally achieve the goal of completely replacing the existing IPv4 network.

REFERENCE

[1] Xie Jianping, HUANG Changfu. Current Situation and Development of Decimal Network [J]. Information Technology and Standardization,2004(04):5-9.

[2] Huang Changfu. Interpretation of Digital Domain Name Specification [J]. Information Technology and Standardization,2006(09):34-36.

[3] Zhang Yunghui, Jiang Xinhua, Lin Zhangxi. Comparison between IPv6 and IPv9 [J]. Computer Engineering,2006(04):116-118.

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

[5] Jung H., & Koh S.J. Mobile Optimized Future Internet", http://protocol.knu.ac.kr/tech/CPL-TR- 10-01-MOFI-12.pdf, Jun. 2010.

Review of Anomaly Detection Based on Log Analysis

Wu Xudong
Laboratory of Wireless Network and Intelligent System
Xi'an Technological University
Xi'an, 710021, China
E-mail: wuxudong_wxd@163.com

***Abstract*—The development of the Internet and the emergence of large-scale systems promote the rapid development of society, and bring a lot of convenience to people. Then comes the problem of network security, privacy theft, malicious attacks and other illegal acts still exist, a qualified software system will log the key operation behavior of the software. Therefore, log analysis has become an important means of anomaly detection. Based on log analysis, this paper consulted the related literature on anomaly detection, elaborated the research status of anomaly detection based on log analysis from the aspects of template matching, rule self-generation and outlier analysis, and analyzed the challenges faced by anomaly detection based on log analysis.**

Keywords-Log Analysis; Distributed; Big Data; Anomaly Detection

I. INTRODUCTION

With the development of the Internet, big data and artificial intelligence have penetrated into people's lives, unknowingly changing the way people live, food, and transportation, making people's lives faster, more efficient, and easier. Research in various fields of computer is moving towards bionics, including human-like big data processing, human-like computer vision and image processing, human-like voice input, etc. These studies make the computer in a domain not only Clairvoyance, Shunfeng ear, can also save and process a large number of various types of data obtained from various aspects, forming an invisible "superman" individual.

In the past 20 years, with the rapid development of the Internet in China, people's lifestyles have undergone tremendous changes. Chinese Internet users continue to grow. According to CNNIC’s 44th "Statistical Report on Internet Development in China", as of June 2019, the number of Internet users in China reached 854 million, an increase of 25.98 million from the end of 2018, and the Internet penetration rate reached 61.2%, compared with the end of 2018. An increase of 1.6 percentage points. The proportions of using desktop computers, laptop computers and tablet computers to surf the Internet were 46.2%, 36.1% and 28.3% respectively. These not only reflect the continuous increase in the number of netizens, but also the rapid and continuous growth of log data from the side.

The log records the time point selected by the developer that is worthy of attention and the changes of state or event that is worthy of attention at this point in time. It is the most important source of information for understanding the operating status of the system and diagnosing system problems. Traditionally, system maintainers use tools such as grep and awk to filter keywords such as "error" or

DOI: 10.21307/ijanmc-2020-036

"exception" in the log to find problems in system operation. When the filtering keywords cannot meet the demand, more experienced personnel will write scripts to impose more complex filtering rules. The cost of this method is very high, writing effective scripts requires a deep understanding of the target system, and these scripts written for specific target systems cannot be applied to other systems, and their versatility is poor. But even without considering the cost, this approach has become no longer feasible for today's software systems.

The ever-increasing log data scale and network security issues make network managers face severe challenges: not only need to ensure the stable and efficient operation of the network, but also need to provide secure network services as much as possible. Fortunately, in recent years, distributed computing technology has become more mature. Distributed computing platforms such as Hadoop, Spark, Flume, Storm are being accepted and applied by more and more companies, and are gradually being used in various industries for data storage. And online or offline analysis, which brings opportunities for log data anomaly detection.

At the same time, issues such as security and privacy in the network have also emerged. Distributed denial of service attacks, zombie codes, Trojan horses, ransomware, worms and other malicious software have a great negative impact on people’s lives. Once the malware operates, it may cause irreversible losses to the company’s economy. , Poses a great threat to people's privacy. A study showed that [1][2]: Random sample surveys of large-scale systems, more than half of the system failure problems were not logged. At this time, maintenance personnel are required to manually find the cause of the problem. Due to the large amount of code, The time invested is much more. High-quality software code can greatly help the detection efficiency after a program error occurs. Log records at key locations are an important means to ensure that the abnormality can be quickly located and repaired. Therefore, it is necessary to add log records to key positions of the program, and log analysis has become an important method of anomaly detection.

The Internet brings convenience to our life, but also brings a series of network security problems. The main characteristics of Internet security problems are as follows: a variety of types, all the time, causing huge losses. All kinds of human attacks, mis-operation, network equipment failure will bring network security problems. Distributed denial of service attack, zombie code, Trojan horse, blackmail program, worm virus and other malicious software appear frequently. Once the malicious software operates, it may cause irreparable loss to the company's economy and cause great impact on people's life.

The log records the time points that developers choose to pay attention to and the changes of states or events at this time point. It is the most important information source to understand the system operation status and diagnose system problems. Traditionally, system maintenance personnel use grep, awk and other tools to filter keywords in logs, such as "error" or "exception", to find problems in system operation. When the filtering keywords can't meet the requirements, senior personnel will write scripts to impose more complex filtering rules. The cost of this method is very high, writing effective scripts needs to have a deep understanding of the target system, but these scripts written for specific target system can not be applied to other systems, and the generality is very poor. But even without considering the cost, this approach is no longer feasible for today's software systems.

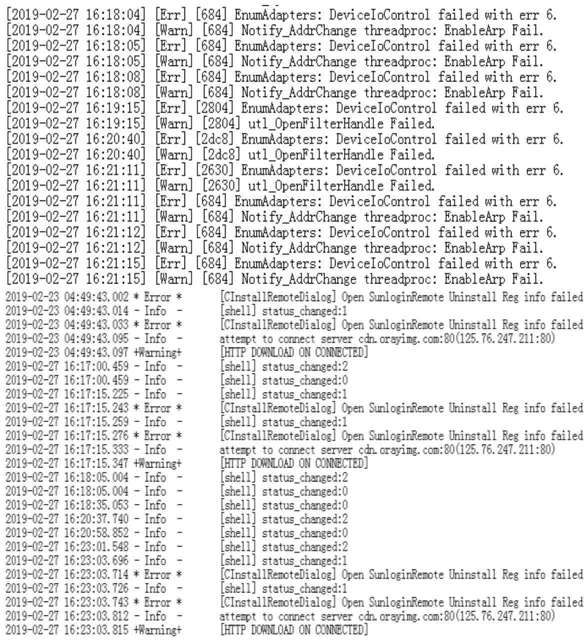
With the increasing scale of log data and network security issues, network managers are facing severe challenges: not only need to ensure the stable and efficient operation of the network, but also need to provide as much as possible secure network services. Fortunately, in recent years, the distributed computing technology is becoming more and more mature. Hadoop, spark, flume, storm and other distributed computing platforms are being accepted and applied by more and more enterprises, and are gradually applied to various industries for data storage and online or offline analysis, which brings opportunities for log data anomaly detection.

This article first talks about the related knowledge of log analysis and anomaly detection, and then summarizes the current research status of log anomaly detection from the aspects of template matching, rule generation and outlier analysis, analyzes and classifies the articles that have been read, and summarizes the current The types and rules of log anomaly detection are found to be difficult to solve during the detection process. Finally, the future work of anomaly detection based on log analysis is summarized.

II. RELATED TECHNOLOGIES AND CONCEPTS OF LOG ANOMALY DETECTION

A. *Log analysis*

The log in the computer is a record of events generated with the operation of network equipment, applications, and systems. Each line records the date, time, type, operator, and description of related operations. Figure 1 shows a partial log record of the application. In reality, the log data generated by a system is very large, conforming to the 4V characteristics defined in big data, namely, volume, variety, velocity, and value. These log data will only occupy storage space if they are shelved, and will bring unlimited value if they are properly used. Because these log data have 4V characteristics, it also determines that manual analysis of these data is unrealistic, and log analysis tools must be used to make full use of the value of log data.



```
[2019-02-27 16:18:04] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:04] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:18:05] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:05] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:18:08] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:08] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:19:15] [Err] [2804] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:19:15] [Warn] [2804] utl_OpenFilterHandle Failed.
[2019-02-27 16:20:40] [Err] [2dc8] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:20:40] [Warn] [2dc8] utl_OpenFilterHandle Failed.
[2019-02-27 16:21:11] [Err] [2630] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:11] [Warn] [2630] utl_OpenFilterHandle Failed.
[2019-02-27 16:21:11] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:11] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:21:12] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:12] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:21:15] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:15] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
2019-02-23 04:49:43.002 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-23 04:49:43.014 - Info  -      [shell] status_changed:1
2019-02-23 04:49:43.033 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-23 04:49:43.095 - Info  -      attempt to connect server cdn.orayimg.com:80(125.76.247.211:80)
2019-02-23 04:49:43.097 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
2019-02-27 16:17:00.459 - Info  -      [shell] status_changed:2
2019-02-27 16:17:00.459 - Info  -      [shell] status_changed:0
2019-02-27 16:17:15.225 - Info  -      [shell] status_changed:1
2019-02-27 16:17:15.243 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:17:15.259 - Info  -      [shell] status_changed:1
2019-02-27 16:17:15.276 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:17:15.333 - Info  -      attempt to connect server cdn.orayimg.com:80(125.76.247.211:80)
2019-02-27 16:17:15.347 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
2019-02-27 16:18:05.004 - Info  -      [shell] status_changed:2
2019-02-27 16:18:05.004 - Info  -      [shell] status_changed:0
2019-02-27 16:18:35.053 - Info  -      [shell] status_changed:0
2019-02-27 16:20:37.740 - Info  -      [shell] status_changed:2
2019-02-27 16:20:58.852 - Info  -      [shell] status_changed:0
2019-02-27 16:23:01.548 - Info  -      [shell] status_changed:2
2019-02-27 16:23:03.696 - Info  -      [shell] status_changed:1
2019-02-27 16:23:03.714 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:23:03.726 - Info  -      [shell] status_changed:1
2019-02-27 16:23:03.743 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:23:03.812 - Info  -      attempt to connect server cdn.orayimg.com:80(125.76.247.211:80)
2019-02-27 16:23:03.815 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
```

Figure 1. Part of the application log

Here are several current mainstream log analysis tools. Slunk is a full-text search engine for machine data and a hosted log management tool. Its main functions include: log aggregation, search, meaning extraction, grouping, formatting, and visualization of results. ELK is composed of three parts: elasticsearch, logstash, and kibana. Elasticsearch is a near real-time search platform. Compared with MongoDB, elasticsearch has more comprehensive functions and is very capable of performing full-text search. It can index, search, and sort documents, filter. Logstash is a log collection tool, which can collect various messages from local, network and other places and send them to elasticsearch. Kibana provides a visual interface on the web and has a cool dashboard.

B. *Store log data*

Due to the huge amount of log data and semi-structured data, the traditional structured database can not meet the storage requirements of log data. HDFS (Hadoop distributed file system) can provide high-throughput data access, which is very suitable for large-scale data sets, and it is suitable for deployment on low-cost machines, which can meet the storage requirements of log data. In the experiment, the log data generated by the system needs to be stored in HDFS. The configured HDFS will automatically back up the data. The input data file is divided into fixed size blocks. The general size of the data block is 128MB. Each data block is stored in different nodes. Generally, each data block has three copies. The first copy is stored in the same node as the client, the second replica exists on a node in a different rack, and the third replica exists on another node in the same rack as the second replica.

C. Log data preprocessing

Log data preprocessing has three goals:

- filtering "non-conforming" data and cleaning meaningless data;
- format conversion and regularization;
- filtering and separating various basic data with different needs according to the subsequent statistical requirements.

In terms of filtering "non-conforming" data and cleaning meaningless data, the log data generated by the system may be "non-conforming" or meaningless. Before the data format conversion and normalization, a judgment needs to be added to check whether the data is standard and intentional. If not, the data is considered useless and jumps to the next data directly. In terms of format conversion and regularization, we first analyze the characteristics of the data. The fields in each record are separated in the form of spaces. According to this feature, each record is segmented according to the space as the standard. For the fields with spaces inside, we need to use regular matching for special processing. After segmentation, each field is normalized, including time format conversion, number type conversion, path completion, etc. In the aspect of filtering and separating data with different needs, the required fields are extracted according to the needs of subsequent detection algorithms.

D. Anomaly detection

Anomalies usually include outliers, fluctuation points and abnormal event sequences. Generally, given the input time series X, the outliers are timestamp value pairs (t, Xt), where the observed value xt is different from the expected value of the time series, then the observed value Xt is an outlier. Fluctuation point refers to a given input time series X, at a certain time t its state or behavior in this time series is different from the values before and after T. An abnormal time series is a part of a given set of time series X={Xi} that belongs to X but is inconsistent with most time series values on X. The abnormal point is given in the box in Figure 2.

Peng Dong et al. [3] divided anomaly detection methods into three categories: techniques based on statistical models, techniques based on proximity, and techniques based on density.

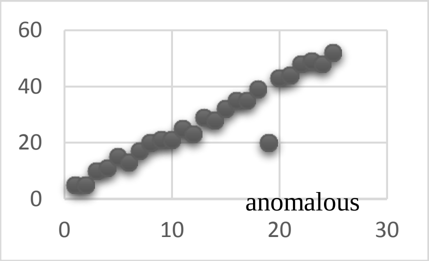


Figure 2. Outlier feature

In data mining, anomaly detection identifies items, events, or observations that do not match the expected pattern or other items in the dataset. Usually, abnormal items will turn into bank fraud, structural defects, medical problems, text errors and other types of problems. Anomalies are also known as outliers, novelty, noise, bias, and exceptions.

Especially in the detection of abuse and network intrusion, interesting objects are often not rare objects, but they are unexpected activities. This pattern does not follow the usual statistical definition of outliers as rare objects, so many anomaly detection methods will fail to deal with such data unless appropriate aggregation is carried out. On the contrary, clustering analysis algorithm may be able to detect the micro clustering formed by these patterns.

There are three types of anomaly detection methods. Under the assumption that most instances in the dataset are normal, the unsupervised anomaly detection method can detect the unlabeled test data by finding the most unmatched instance with other data. Supervised anomaly detection requires a dataset that has been labeled "normal" and "abnormal" and involves training classifiers (the key difference from many other statistical classification problems is the inherent imbalance of anomaly detection). The semi supervised anomaly detection method creates a model representing normal behavior based on a given normal training data set, and then detects the possibility of test cases generated by the learning model.

III. RESEARCH STATUS OF LOG ANOMALY DETECTION TECHNOLOGY

Anomaly detection refers to the process of finding data patterns that do not meet expectations from the data [4]. Anomaly detection behavior based on log data can be regarded as a classification problem in essence, that is to say, distinguish normal behavior and abnormal behavior from a large amount of abnormal log behavior data, and determine the specific attack method from the abnormal behavior [5]. When the server is running, the log will record and generate the behavior of the user throughout the access process. You can find the information of abnormal users by processing the information in the log. Therefore, analyzing logs has become one of the most effective methods to detect abnormal user behavior [6,7,8]. With the rapid development of big data, log-based anomaly detection methods are divided into three categories: model-based technology, proximity-based technology and density-based technology.

A. Model-based technology

Model-based technology first builds a data model. Anomalies are objects that the model cannot fit perfectly. Since abnormal and normal objects can be regarded as defining two different classes, we can use classification techniques to build models of the two classes. However, the training set is very important in the classification technology. Because anomalies are relatively rare, it is impossible to detect new anomalies that may appear [9]. Wang Zhiyuan et al. [10] used the log template to detect anomalies in 2018. The log was cleaned first, and then the edit distance was used to cluster the text to form the log template. On the basis of the log template, TF-IDF (Word Frequency-Inverse File) was used. Frequency) to form a feature vector, and then use logistic regression, Bayesian, support vector machine and other weak classifiers to train to obtain the score feature vector, build a strong classifier based on the score feature vector and random forest, and finally use mutual information to detect the truth The correlation between the template and the clustering template, the accuracy and recall rate are used to detect the classification effect and various classifiers are compared. Siwoon et al. [11] proposed a new data storage and analysis architecture based on Apache Hive to process a large amount of Hadoop log data, using average movement and 3-sigma technology to design and implement three anomaly detection methods, these three methods They are the basic method, linear weight method and exponential weight method. The first method calculates the average line and standard deviation of anomaly detection, but there are repeated detections. In order to solve this problem, there are two other weighted detection methods, namely linear weighting and exponential weighting. In the linear weighting method, the weight is given in proportion to the position of the log item, and the exponential weight method is to give the weight exponentially on top of the basic method. Finally, the effectiveness of the proposed method is evaluated in a hadoop environment with a name node and four data nodes. Fu et al. [12] proposed a technique that does not require any specific application knowledge for anomaly detection in unstructured system logs, including a method of extracting log keys from free text messages. The false positive rate of their experiments under the Hadoop platform is about 13%. Xu et al. [13] used source code to match the log format to find the relevant variables, extracted the features of the corresponding log variables through the bag-of-words model, and then used these features to reduce the dimensionality through the principal component analysis method, according to the maximum separability of principal component analysis Detect abnormal log files, and finally use a decision tree to visualize the results. Fronza et al. [14] used the operation sequence represented by the random index, characterized the operation in each log according to its context, and then used the support vector machine to correlate the sequence to the fault or non-fault category to predict system failure . Peng et al. [15] applied text mining technology to classify messages in log files as common cases, improved classification accuracy by considering the time characteristics of log messages, and used visualization tools to evaluate and verify the effective time for system management mode.

B. *Technology based on proximity*

Proximity-based technology considers proximity measures between objects, such as "distance". Zhang Luqing et al. [16] proposed a web attack data mining algorithm based on the anomaly degree of outliers, which first clustered HTTP requests, and then proposed a detection model that approximates normal distribution. The algorithm first finds the arithmetic mean of each numerical attribute value and the most frequently occurring value in each categorical attribute as the centroid of the numerical attribute and the centroid of the categorical attribute, and after synthesis, the centroid T of the data set is obtained, and the distance between the object p and the centroid T is obtained. As the abnormality of p. Finally, experiments have confirmed that the algorithm has a higher detection rate. Jakub Breier et al. [17] proposed a log file anomaly detection method, which dynamically generates rules from certain patterns in sample files and can learn new types of attacks while minimizing the need for human behavior. The implementation uses the Apache Hadoop framework to provide distributed storage and distributed data processing to support parallel processing to speed up execution. Since the incremental mining algorithm based on the local outlier factor requires multiple scans of the data set, Zhang Zhongping et al. [18] proposed a flow data outlier mining algorithm (SOMRNN) based on inverse k nearest neighbors. Using the sliding window model to update the current window requires only one scan, which improves the efficiency of the algorithm. Grace et al. [19] used data mining methods to analyze Web log files to obtain more information about users. In their work, they describe the log file format, type and content, and provide an overview of the Web usage mining process. Liang Bao et al. [20] proposed a general method for mining console logs to detect system problems. First give some formal problem definitions, and then extract a set of log statements in the source code and generate a reachability graph to show the reachability of log statements. After that, log files are analyzed to create log messages by combining information about log statements with information retrieval techniques. The grouping of these messages is tracked according to the execution unit. A detection algorithm based on probabilistic suffix tree is proposed to organize and distinguish the significant statistical characteristics of sequences. Experiments were conducted on the CloudStack test platform and Hadoop production system, and the results showed that compared with the existing four algorithms for detecting abnormalities, this algorithm can effectively detect abnormal operation. Since there are fewer abnormal points in reality, Liu et al. [21] proposed an anomaly detection algorithm based on isolation. The isolation tree created can quickly converge but requires sub-sampling to achieve high accuracy.

C. Density-based technology

Density-based technology considers objects in low-density areas as abnormal points. The density-based local outlier detection algorithm (LOF) has high time complexity and is not suitable for outlier detection of large-scale data sets and high-dimensional data sets. Wang Jinghua et al. [22] proposed a local outlier Point detection algorithm NLOF. Li Shaobo et al. [23] proposed a density-based abnormal data detection algorithm GSWCLOF. The algorithm introduces the concept of sliding time window and grid. In the sliding time window, the grid is used to subdivide the data, and the information entropy is used for all The data in the grid is pruned and filtered to eliminate most of the normal data, and finally the outlier factor is used to make a final judgment on the remaining data. Wang Qian et al. [24] proposed a density-based detection algorithm, which introduced the Local Outlier Factor (LOF), and judged whether the data is abnormal based on the LOF value of the data. The algorithm is only suitable for static data detection. Once the amount of data fluctuates, it is necessary to recalculate the LOF value of all data. The algorithm has poor adaptability and is not suitable for the detection process of dynamic data. Pukelsheim et al. [25] assumed that the data sample obeys a univariate Gaussian distribution, and judged the test sample that is outside of the distance twice or three times the variance as abnormal.

IV. CHALLENGES FACED BY LOG ANOMALY DETECTION TECHNOLOGY

There are several obstacles from the time the system abnormality occurs to the successful detection of the abnormality:

- The exception log is not recorded
- The format of exception log records is not standardized
- The exception log cannot be sent to the processing end in time
- Abnormal log sending is lost
- The detection algorithm is not accurate enough

Any occurrence of one or more of the above conditions will result in failure of the anomaly detection result.

A. Real-time

The purpose of anomaly detection is to find anomalies and find a corresponding method to float the anomaly, and the time delay from logging, to anomaly detection, to manual analysis, and to anomaly elimination is too long, which leads to anomalies that exist for too long. The losses were more serious. If real-time performance can be guaranteed, the efficiency of exception elimination will be greatly improved.

B. Detection accuracy

Anomaly detection has various factors that affect its accuracy, such as irregular log format, inappropriate algorithm, etc. These problems directly lead to a decrease in the accuracy of anomaly detection, which also determines that log anomaly detection cannot be completely separated from the intervention of technicians.

Even if the same benchmark data set is used in the literature for anomaly detection, most of them do not indicate the size or proportion of labeled data. Even the size of training and test sets and evaluation indicators are different. Different measurement combinations make the research results unable to compare with each other

C. The versatility of detection algorithms

At present, there are many anomaly detection algorithms at home and abroad, such as: Isolation Forest, One-Class SVM, Robust covariance, K-means, Principal Component Analysis, 3-ε, etc. These algorithms have their own advantages and disadvantages and are not suitable for all anomaly detection. However, due to its unstructured and non-identical characteristics of logs, a specific algorithm is needed for a specific log, or a specific algorithm is improved to achieve a higher detection rate. The "localization" of the algorithm also requires specialized technical personnel to operate, which increases the cost of detection.

D. Tag data

In the log data, there is a large amount of data, and there are very few abnormal data. It is very difficult to mark a small amount of abnormal data in a large amount of data. There is no such publicly marked data as the experimental basis, so anomaly detection encountered great difficulties.

V. RESEARCH DIRECTION OF ANOMALY DETECTION

Based on the current research status of anomaly detection technology and the above problems, the challenges and future research directions of anomaly detection are summarized as follows:

Traffic data often have high characteristic dimensions, and the Euclidean distance in the sampling method can not measure the spatial distribution of the samples very well. The data distribution environment of supervised learning and semi supervised learning are different. Under unbalanced data, most of the existing semi supervised methods apply the traditional methods to semi supervised learning. Therefore, the traditional methods to solve the imbalance problem are not necessarily suitable for semi supervised learning and need further research. Although the research on data imbalance has achieved good results in the field of network security, there are very few researches on the imbalance problem in semi supervised learning. Most of the semi supervised methods applied in the field of anomaly detection use ensemble learning to solve the class imbalance. In the future, we can solve the problem of anomaly detection by combining the latest achievements in the field of data imbalance under semi supervision.

At present, many network traffic feature selection and extraction are limited to one dimensional features or simple combination of multi-dimensional features, while traffic anomalies usually show in multi-dimensional features. How to effectively fuse multi-dimensional features, learn data flow features from multiple perspectives, and use a small amount of labeled data for semi supervised integration algorithm synthesis results to reduce information loss is a challenging research topic.

Semi supervised dimensionality reduction is a feasible method in the field of anomaly detection. How to find a more effective way to deal with high-dimensional sparse samples and continuous variables and further improve the real-time performance of detection model is of great significance.

The learning effect of the combination of active learning and semi supervised learning strategy is better than that of single method. The combination of semi supervised learning and active learning can actively find effective supervision information. Through effective supervision information, unlabeled sample data can be used better, thus improving the accuracy of the model and solving speed. However, the research on the combination of semi supervised learning and active learning is rare, and there is a large space for improvement.

Incremental semi supervised anomaly detection is more in line with the actual anomaly detection. It makes full use of the data results processed before in the training process. It should have more in-depth research in the field of network security. In the future, we can consider introducing the incremental algorithm of natural language technology into specific anomaly detection.

Semi supervised clustering algorithm uses the traditional clustering algorithm to introduce the supervised information to complete the semi supervised learning, so it can also expand the semi supervised clustering algorithm such as density clustering and spectral clustering. In addition, some traffic data are high-dimensional and sparse. However, most of the existing clustering algorithms are not suitable for processing high-dimensional sparse data. In future research, it is necessary to make further discussion.

In general, semi supervised learning can help improve performance by using unlabeled data, especially when the number of labeled data is limited. However, in some cases, the selection of unreliable unlabeled data may mislead the formation of classification boundaries and eventually lead to the degradation of semi supervised learning performance. Therefore, how to use unlabeled data safely is a research focus in the future.

It can combine multiple semi supervised anomaly detection methods and technologies to achieve more efficient network data detection and obtain more accurate prediction results. In addition, in semi supervised anomaly detection, it is a challenging research topic to minimize the additional impact on the network.

VI. CONCLUSION

Machine learning faces many challenges in the field of abnormal traffic detection. The biggest difficulty is the lack of label data. In practice, only a limited number of tagged data is available, while most of the data is unmarked. In addition, although there are a large number of normal access data, there are few abnormal traffic samples and various attack forms, which make it difficult to learn and train the model. Semi supervised learning is an effective solution, which can make use of both unlabeled data and labeled data, which can alleviate this problem.

For anomaly detection based on log analysis, domestic and foreign countries have made certain progress and achieved various results. Various algorithms such as template matching, automatic rule generation, outlier analysis, and statistical data have certain effects, which are of great significance to network security and intelligent operation and maintenance.

Future research will continue to focus on real-time performance to ensure that abnormalities can be detected as quickly as possible. Improve detection accuracy, minimize manual intervention or cancel manual intervention. Study the versatility of the algorithm, so that an algorithm can adapt to log analysis in different environments as much as possible.

REFERENCES

[1] Yuan D, Park S, Huang P, Liu Y, Lee MM, Tang X, Zhou Y, Savage S. Be conservative: enhancing failure diagnosis with proactive logging. In: Proc. of the 10th Symp. on Operating Systems Design and Implementation (OSDI). 2012. 293~306.

[2] Yuan D, Park S, Zhou Y. Characterizing logging practices in open-source software. In: Proc. of the 2012 Int'l Conf. on Software Engineering. 2012. 102~112. [doi: 10.1109/ICSE. 2012.6227202].

[3] Peng Dong. Intelligent operation and maintenance: building a large-scale distributed AIOps system from zero. Electronic Industry Press, 2018.7 ISBN 978-7-121-34663-7 p198-p199.

[4] Varun Chandola, Arindam Banerjee, Vipin Kumar. Anomaly Detection: A Survey[J]. Acm Computing Surveys, 2009, 41(3).

[5] Davis J J, Clark A J. Data preprocessing for anomaly based network intrusion detection: A review[J]. Computers & Security, 2011, 30(6-7):353-375.

[6] Q. Lin, H. Zhang, J. Lou, Y. Zhang and X. Chen, "Log Clustering Based Problem Identification for Online Service Systems," 2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C), Austin, TX, 2016, pp. 102-111.

[7] Pecchia A, Cotroneo D, Kalbarczyk Z, et al. Improving Log-based Field Failure Data Analysis of multi-node computing systems[C]. IEEE, 2011.

[8] Tambe R, Karabatis G, Janeja V P. Context aware discovery in web data through anomaly detection[J]. International Journal of Web Engineering and Technology, 2015, 10(1):3.

[9] Wang Xiaodong, Zhao Yining, Xiao Haili, Chi Xuebin, Wang Xiaoning. Detection method of abnormal log flow pattern in multi-node system [J/OL]. Journal of Software: 1-15 [2019-12-24].

[10] Wang Zhiyuan, Ren Chongguang, Chen Rong, Qin Li. Anomaly detection technology based on log template[J]. Intelligent Computers and Applications, 2018, 8(05): 17-20+24.

[11] Son S, Gil MS, Moon YS. [IEEE 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)-Jeju Island, South Korea (2017.2.13-2017.2.16)] 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)-Anomaly detection for big log data using a Hadoop ecosystem[J]. 2017:377-380.

[12] Fu, Q., Lou, JG, Wang, Y., & Li, J. (2009). Execution anomaly detection in distributed systems through unstructured log analysis. In Proceedings of the 2009 ninth IEEE international conference on data mining, ICDM '09, (pp. 149–158). Washington, DC: IEEE Computer Society. doi:10.1109/ICDM. 2009.60.

[13] Xu W, et al. Large-scale system problems detection by mining console logs[J]. Proceedings of the Acm Sigops Symposium on Operating Systems Principles Big Sky Mt, 2013:2009.

[14] Ilenia Fronza, Alberto Sillitti, Giancarlo Succi, Mikko Terho, Jelena Vlasenko. Failure prediction based on log files using Random Indexing and Support Vector Machines[J]. Journal of Systems and Software, 2013, 86(1):2- 11.

[15] Peng W, Li T, Ma S. Mining logs files for data-driven system management. ACM SIGKDD Explorations Newsletter, 2005, 7(1):44-51.

[16] Zhang Luqing. Web attack data mining algorithm based on outlier anomaly[J]. Ship Electronic Engineering, 2018, 38(09): 105-110.

[17] Breier J, Jana Branišová. A Dynamic Rule Creation Based Anomaly Detection Method for Identifying Security Breaches in Log Records[J]. Wireless Personal Communications, 2015, 94(3):1-15.

[18] Zhang Zhongping, Liang Yongxin. Algorithm for mining outliers in flow data based on anti-k nearest neighbors[J]. Computer Engineering, 2009, 35(12): 11-13.

[19] Grace, L., Maheswari, V., & Nagamalai, D. (2011). Web log data analysis and mining. In N. Meghanathan, B. Kaushik, & D. Nagamalai (Eds.), Advanced computing, communications in computer and information science (Vol. 133, pp. 459–469). Berlin: Springer.

[20] Liang Bao, Qian Li, Peiyao Lu, Jie Lu, Tongxiao Ruan, Ke Zhang. (2018). Execution anomaly detection in large-scale systems through console log analysis. The Journal of Systems & Software 143 (2018) 172– 186.

[21] Liu F T, Ting K M, Zhou Z H. Isolation-Based Anomaly Detection[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1):1-39.

[22] Wang Jinghua, Zhao Xinxiang, Zhang Guoyan, Liu Jianyin. NLOF: A new density-based local outlier detection algorithm [J]. Computer Science, 2013, 40(08): 181-185.

[23] Li Shaobo, Meng Wei, Wei Jinglei. Density-based abnormal data detection algorithm GSWCLOF[J]. Computer Engineering and Applications, 2016, 52(19): 7-11.

[24] Wang Qian, Liu Shuzhi. Improvement of local outlier data mining method based on density [J]. Application Research of Computers, 2014, 31(06): 1693-1696+1701.

[25] Pukelsheim F. The Three Sigma Rule[J]. The American Statistician, 1994, 48(2):88-91.

Software TLB Management Method Based on Balanced Binary Tree

Chen Hongyu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 15771900781@189.cn

Zhao Li
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zhaoli1998@163.com

Zhang Yuke
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 17792012345@189.cn

Ai Jian
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: aijianup@163.com

***Abstract*—Purpose: with the development of the computer system, there is a higher request to reduce the failure times of Translation Look—aside Buffer and relieve the failure influence, normal solution is deal with TLB failure by software or hardware, it find the page table and then implement the index operation to locate the page want. Method: In order to satisfy the needs of software management method for mapping speed from virtual address to physical address and to expand the size of TLB, our team design a software management method based on balanced binary search tree. Page management is implemented in software management, TLB is managed by operating system based on abstract model, and MMU (memory management unit) is no longer used Unit, our team build a balanced binary tree to search TLB. Result: Binary search tree has the advantage of pruning in the interval search problem under the balanced state. Therefore, searching TLB by this method is conducive to the expansion of TLB capacity, making space for cache and other performance improvement designs on chip and reducing costs. However, the search speed is reduced and some time is sacrificed to free up CPU space. Conclusion: This method can free up CPU design space, also can expand more size of TLB and reduce the cost, its optimization algorithm is worthy of further research.**

Keywords-Component; TLB; Balance Binary Tree; Management of Software

I. INTRODUCTION

With the continuous development of the computer, the speed of CPU is faster and faster, but the speed of memory has not been improved. The development of TLB will help computer to process large virtual address. In this paper, our team design a TLB software management method of constructing a balanced binary search tree based on virtual page number. When the TLB capacity of balanced binary search tree is large, the method could deal with it quickly, for example, the number of virtual page numbers in balanced state is doubled, and the average retrieval times only need to be increased once.

Therefore, the retrieval speed is fast, the capacity increases, and the hit rate will also increase. Moreover, the time cost of tree building is evenly distributed in each search, and the average search time decreases. The balanced binary search tree is used to cut the interval search the advantage of branch and search is to quickly search the virtual page number and locate it to the corresponding location of the memory. Because the page number needs to be searched many times, it helps to deal with the large virtual space.

II. TLB

Most programs always visit a few pages many times, so TLB records frequent pages and their information, which can accelerate the mapping from virtual address to physical address.

DOI: 10.21307/ijanmc-2020-037

The basic unit of TLB internal storage is the page table item, which corresponds to the page table item stored in RAM, the more TLB capacity, the more page table items can be stored, and increase the probability of TLB hit rate. Due to the limited capacity of TLB, RAM page table and TLB page table items cannot be one-to-one correspondence.

A. *Location*

TLB is used to cache some tab table entries. TLB can be between CPU and CPU cache, or between CPU cache and main memory, depending on whether cache uses physical addressing or virtual addressing. If the cache is a virtual addressing, the addressing request is sent directly from the CPU to the cache, and then the required TLB entries are accessed from the cache. If the cache uses physical addressing, the CPU will first perform for each memory operation and send the obtained physical address to the cache. Each method has its own advantages and disadvantages.

B. *Common optimization*

A common optimization of cache with physical addressing is parallel TLB search and cache access. The lower bits of all virtual addresses (e.g., the lower 12 bits in the virtual address when there is a 4KB tab in the virtual memory system) represent the address offset (in page address) of the requested address within the paging, and these bits will not change during the transition from the virtual address to the physical address. The process of accessing the CPU cache consists of two steps: use an index to find the corresponding entries in the CPU cache data store, and then compare the corresponding tags of the CPU cache entries found. If the cache is indexed by the same page address during the translation of addresses, the translation of higher bits of virtual and real address (i.e. page to page address / page number of pages) on TLB and the "index" operation of CPU cache can be performed in parallel. The page number of the physical address obtained from the TLB is then sent to the CPU cache. The CPU cache compares page number tags to determine whether the access is missing or missing. It is also possible to perform TLB search and CPU cache access in parallel, even if the CPU cache must use some bits that may change after address translation. Refer to the address translation section of cache entry for further details on cache and TLB under virtual addressing.

III. ALGORITHM OF BALANCED BINARY SEARCH TREE

Balanced binary tree has the following properties: it is an empty tree or the absolute value of the height difference between its left and right sub trees is no more than 1, and the left and right sub trees are both balanced binary tree.

A. *Insert operation*

Inserting a new node into the balanced binary tree destroys the balance of the balanced binary tree. First of all, our team need to find the pointer of the root node of the minimum sub tree which is out of balance after inserting a new node. Then adjust the link relationship between the nodes in the sub tree to make it a new balanced sub tree. When the unbalanced minimum sub tree is adjusted to a balanced sub tree, all the other unbalanced sub trees do not need to be adjusted.

LL type adjustment: Insert a node on the left sub tree of point B. After insertion, the balance factor of the left sub tree of point B becomes 1 and that of node a becomes 2. In this way, it can see that the sub tree with node a as the root node is the minimum unbalanced sub tree. When adjusting, the left child B of a is rotated to the right instead of a as the root node of the original unbalanced sub tree, and the lower right rotation of the node of a is called the root node of the right sub tree of B, and the original right sub tree of B becomes the left sub tree of A.

In the binary search tree insertion and deletion operations, the advantage of using balanced tree is to make the tree structure better, so as to improve the speed of search operation. The disadvantage is that the insertion and deletion operations become more complicated, which reduces their operation speed. The operation of the imbalance caused by deleting a node in a binary search tree is more complex than that of inserting a node, so it will not be discussed here.

B. *AVL Deletion*

Like the insert operation, deleting a node may break the balance, which requires us to adjust the balance when deleting.

The deletion is divided into the following situations:

- First, search the whole binary tree for the node to be deleted. If it is not found, it will be returned without processing. Otherwise, the following operations will be performed.The node to be deleted is the current root node t.If the left and right sub trees are not empty. The deletion operation is implemented in the higher sub tree.
- The height of the left sub tree is greater than that of the right sub tree. Assign the largest

element in the left sub tree to the current root node, and then delete the node with the largest element value in the left sub tree.

- If the height of the left sub tree is less than that of the right sub tree, assign the smallest element in the right sub tree to the current root node, and then delete the node with the smallest element value in the right sub tree.
- If one of the left and right sub trees is empty, replace the current root node with the non empty sub tree or null.
- If the element value of the node to be deleted is less than the T value of the current root node, delete it in the left sub tree.
- Recursively, delete in the left sub tree.

This is to determine whether the current root node still meets the equilibrium condition.

If the equilibrium condition is satisfied, only the height information of the current root node T needs to be updated. Otherwise, rotation adjustment is required.

If the height of the left sub tree of the left child node of T is greater than the height of the right sub tree of the left child node of T, the corresponding single rotation is performed. Otherwise, double rotation is performed.

The element value of the node to be deleted is greater than the T value of the current root node. Delete it in the right sub tree.

C. *Summary:*

- Non leaf nodes have at most two child nodes.
- The value of non leaf node is larger than that of left child node and less than that of right child node.
- The difference in the number of levels on the left and right sides of the tree is no more than 1.
- There are no nodes with identical values.

IV. DESIGN SCHEME

The application of page number can be approximately regarded as a combination of mass or partially ordered intervals over time. When the CPU offer page number request, the operating system first searches the balanced binary search tree about page number in TLB, and compares the root node with the request page number. If the page number of the root node is greater than the request page number, the operating system first searches the balanced binary search tree of the TLB, Recursively search in the left sub tree. If the page number of the root node is less than the request page number, then it is recursively searched in the right sub tree. If the page number of the root node is exactly equal to the request page number, the query is considered successful. The corresponding physical block will read out, and the adder is used to splice the address in the page to convert it into the corresponding physical address.

A. *Recursion exit*

Therefore, there are two exits for the end of recursion.

- One is that the access is successful and the corresponding physical address is obtained;
- the other is that the left or right sub tree of the root node is empty. If the access fails, there is no corresponding page number in the TLB. The operating system will retrieve the page table in memory, call the corresponding page table entry into TLB, and insert it into the balanced binary search of the TLB according to the page table entry in the tree.

When the balanced binary tree constructed by TLB is full, it is necessary to replace the oldest page table entries. For deleting a page table item, it is necessary to set different flag bits for replacement operation which accord to the corresponding replacement algorithm. Deletion will also cause imbalance, it is necessary to balance the binary search tree first after deletion.

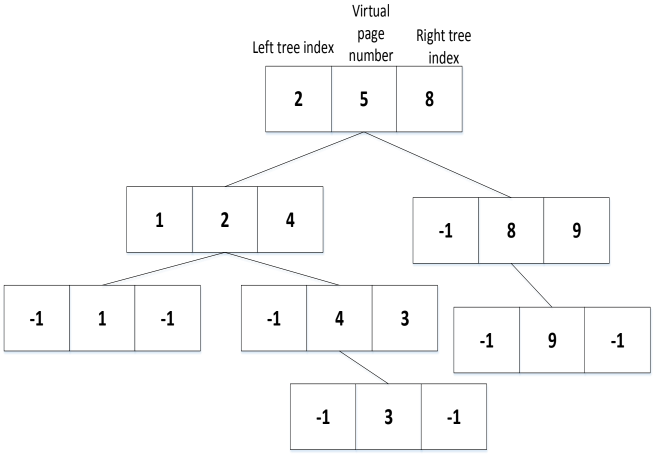


Figure 1. Model

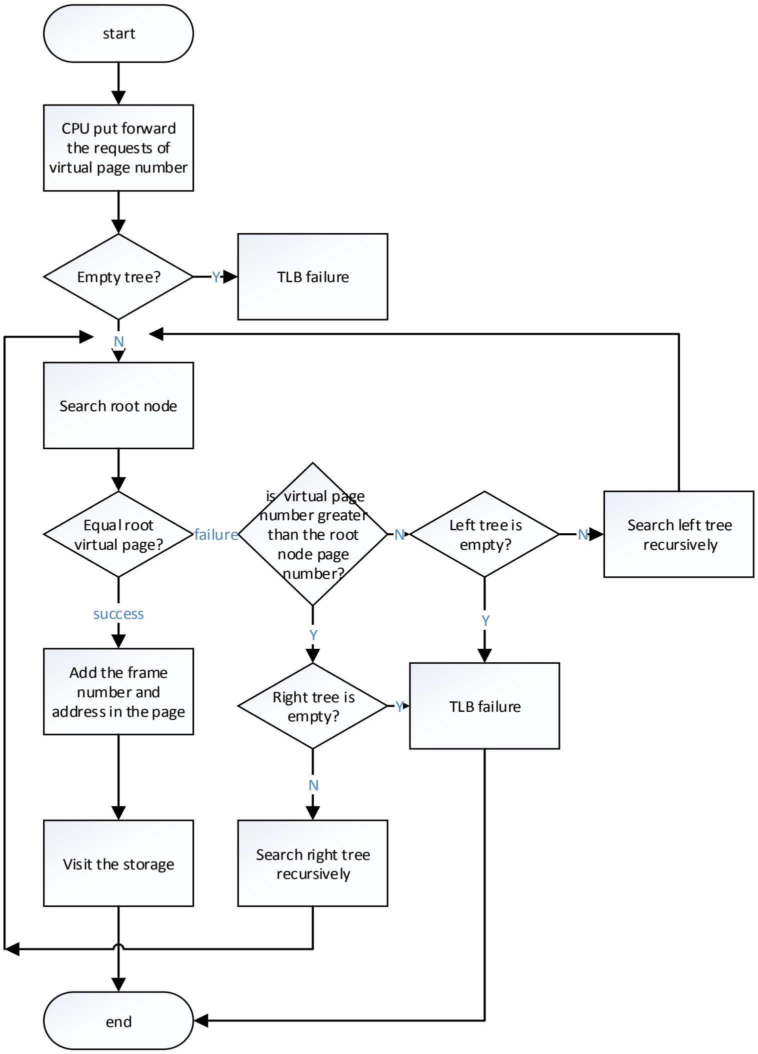

start
CPU put forward the requests of virtual page number
Empty tree?
Y
TLB failure
N
Search root node
Equal root virtual page?
failure
is virtual page number greater than the root node page number?
N
Left tree is empty?
N
Search left tree recursively
success
Y
Y
Add the frame number and address in the page
Right tree is empty?
Y
TLB failure
N
Visit the storage
Search right tree recursively
end

Figure 2. Search step

V. TLB STORAGE

A typical page table entry includes page frame number, protection bit, modification bit (dirty bit), access bit, cache forbidden bit and on / off bit.

TLB has virtual page number, significant bit, modification bit, protection bit, page frame number, root node index bit of left and right sub tree, balance factor and so on.

0	1	2	3	4	19	34	49
Significant	protect	modify	balance	Page frame number	Left tree index	Page number	Right tree number

Figure 3. Item

A. Composition

The virtual page number marks the page corresponding to the table item, and the significant bit records whether the table entry is used or not, whether the table item has been modified, the read / write and execution permissions of the protection bit record, and the setting of index bit and balance factor bit of the root node of the left and right sub trees are determined by comparing the size of the root node with that of the root node to determine whether the search of the left and right sub trees is recursive .

The node stores the index of the left and right child nodes, and then stores the balance factor (the balance factor is only 1,0, -In three cases, if the absolute value exceeds 1, then the balanced binary search tree is judged to be unbalanced).

When the TLB is full, some page table items need to be eliminated. Therefore, according to the corresponding page replacement algorithm, different flag bits are flexibly set to mark the page table items that should be eliminated; the row significant bit is that when the system starts, each TLB row is empty, and the information in it is invalid In order to show whether the information in the TLB row is valid, each row has a significant bit. By clearing the significant bit of the row, the corresponding page table item is eliminated.

Note: setting the global bit in a page directory/table entry will prevent that entry from being flushed. This is useful for pinning interrupt handlers in place.

B. *Alternate method*

An alternate method is to use instruction, which should be used instead of the above method when doing small mapping modifications (creation, removing, changing.) instruction is mostly used in page not mapping and remapping routines in order to invalidate a previous cached translation. If instruction or some other TLB flush method had not been used, the mapping would remain cached, producing undefined consequences.

However, please note that the instruction instruction was introduced in the i486 ISA and is not part of the i386 ISA, thereby requiring a properly written i386-compatible kernel to use conditional inclusion of relevant code at compilation time depending on the target machine. The above is more complicated in the multiprocessor case. If another processor could also be affected by a page table write (because of shared memory, or multiple threads from the same process), it must also flush the TLB on those processors. This will require some form of inter-processor communication.

VI. Mathematical Verification

Time complexity: the difference between height of the left sub tree and the right sub-tree is no more than 1. Our team assumed that NH is the minimum number of nodes in the balanced binary tree with depth H.

From the method of recursion recursion, the result is NH = NH-1 + NH-2 + 1 (1 is the root node), and the minimum search length is deduced.

Therefore, the equation can be obtained by the characteristic equation method.

$$\lambda^2=\lambda+1$$

The result is $\lambda=\frac{1\pm\sqrt{5}}{2}$

Then our team assign the special solution is n = C, replace it into the former formula, and the result is C = - 1, so the general term formula can be obtained.

$$C_1\lambda_1^{\,n}+C2\lambda_2^{\,n}-1$$

Therefore, the equation can be obtained.

$$\left(1+\frac{2}{\sqrt{5}}\right)\left(\frac{1+\sqrt{5}}{2}\right)^n+\left(1-\frac{2}{\sqrt{5}}\right)\left(\frac{1-\sqrt{5}}{2}\right)^n-1$$

And then figure the equation

$$n<\log\frac{1+\sqrt{5}}{2}(N+1)<\frac{3}{2}\log_2(N+1)$$

According to the characteristics of the tree, the maximum depth of the balanced binary tree with n nodes is log2N, and the average search length is log2N approximately.

VII. Modern Operating System

In modern processor, software uses virtual address to access memory, and MMU unit of processor is responsible for converting virtual address to physical address. In order to complete this mapping process, software and hardware jointly maintain a multilevel mapping page table. When the processor finds that the page table cannot be mapped to the corresponding physical address, it will trigger a page missing exception and suspend the error process. The operating system software needs to handle the page missing exception.

A. *Content*

TLB is specifically used to cache page table entries in memory, usually inside the MMU unit. TLB is a very small cache, and the number of TLB entries is relatively small. Each TLB entry contains information about a page, such as significant bit, virtual page number, modified bit, physical page frame number, etc. When the processor wants to access a virtual address, it will first query in the TLB.

- If there is no corresponding entry in the TLB table entry, it is called TLB miss, then need to access the page table to calculate the corresponding physical address. If there is a corresponding entry in the TLB table entry, the physical address is directly obtained from the TLB table entry, which is called TLB hit.

- The basic unit of TLB internal storage is TLB table entries. The larger the TLB capacity, the more TLB entries can be stored, and the higher the TLB hit rate. However, the capacity of TLB is limited. At present, the Linux kernel uses 4KB small pages by default. If a program uses 512 small pages, it needs at least 512 TLB entries to ensure that TLB miss will not occur. However, if a 2MB page is used, only one TLB table entry is needed to ensure that TLB miss will not occur. For large applications that consume memory in gigabytes, large pages in 1GB can also be used to reduce TLB miss.
- Because accessing the page table in memory is relatively time-consuming, especially when multilevel page tables are widely used nowadays, multiple memory accesses are required. In order to speed up the access, the system designer has designed a hardware cache TLB for page table.

The CPU will first look it up in TLB, because it is very fast to find it in TLB. The reason why TLB is fast is that it contains a small number of entries. On the other hand, TLB is integrated into the CPU. It can run almost at the speed of the CPU.

If an entry (TLB hit) containing the virtual address is found in the TLB, the corresponding physical address can be obtained directly from the entry.

Otherwise, unfortunately, TLB miss will occur, and the page table of the current process (paging structure caches may be used here). At this time, another part of the MMU, the table walk unit, is called out. The table in the MMU is page table.

The method of using table walk unit hardware unit to find page table is called hardware TLB miss handling, which is usually adopted by CISC architecture processor (such as IA-32).

It can't be found in page table, and it will be handled by software (operating system) when page fault appears.

In contrast, software TLB miss handling, which is usually adopted by RISC architecture processors (such as alpha), does not involve the CPU after TLB miss. The operating system searches the page table through software. The way to use hardware is faster, while the way to use software is more flexible. IA-64 provides a hybrid mode, which can take into account the advantages of both.

If the P (present) bit of the entry corresponding to the virtual address is found in the page table, it indicates that the physical page corresponding to the virtual address currently resides in memory, that is, page table hit. There are still two things to do.

Since it can't find it in TLB, it's natural to update TLB.

Check the permissions, including read / write / executable permissions, user / supervisor mode permissions, etc. If it do not have the correct permissions, SIGSEGV (segmentation fault) will be triggered.

If the P bit of entry corresponding to the virtual address is 0, page fault will be triggered. There may be several situations.

B. Address

The virtual address has never been accessed after it has been allocated (for example, if there is no space allocated after allocation, the physical memory will not be allocated). After the page fault is triggered, the physical memory is allocated, that is, demand paging. After a certain demand is available, the system will allocate the P position to 1.

The content of the corresponding physical page is swapped out to the external disk / flash. At this time, the page table entry stores the temporary location of the page in the external swap area. It can switch it back to physical memory, establish the mapping again, and then set P position 1.

If the virtual address does not exist in the page table of the process, it means that the virtual address is not in the address space of the process. At this time, segmentation fault will also be triggered. The CPU’s MMU locates the page directory for the process using the special register mentioned above.

The page directory index (from the first 10 bits of the virtual address) is used to locate the PDE that identifies the page table needed to map the virtual address to a physical one. The page table index (from the second 10 bits of the virtual address) is used to locate the PTE that maps the physical location of the virtual memory page referenced by the address.

The PTE is used to locate the physical page. If the virtual page is mapped to a page that is already in physical memory, the PTE will contain the page frame number (PFN) of the page in physical memory that contains the data in question. (Processors reference memory locations by PFN).

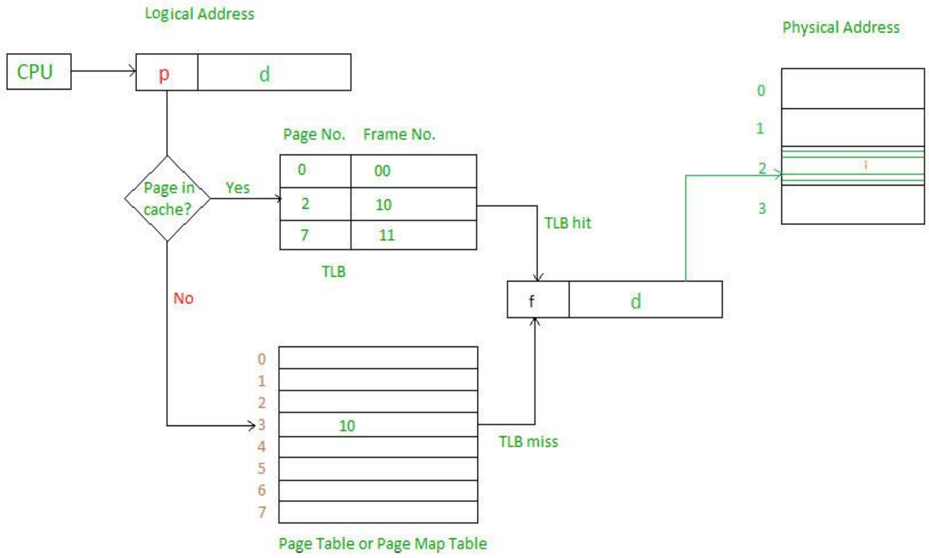


Figure 4. process

C. *Steps in TLB hit:*

- CPU generates virtual address.
- It is checked in TLB (present).
- Corresponding frame number is retrieved, which
- now tells where in the main memory page lies.

D. *Steps in Page miss:*

- CPU generates virtual address.
- It is checked in TLB (not present).
- Now the page number is matched to page table residing in main memory (assuming page table contains all PTE).
- Corresponding frame number is retrieved, which now tells where in the main memory page lies.
- The TLB is updated with new PTE (if space is not there, one of the replacement technique comes into picture i.e either FIFO, LRU or MFU etc).

E. *Conclusion*

Effective memory access time(EMAT) : TLB is used to reduce effective memory access time as it is a high speed associative cache.

EMAT = h*(c+m) + (1-h)*(c+2m)

Where, h = hit ratio of TLB

m = Memory access time

c = TLB access time

If the page is not in physical memory, the MMU raises a page fault, and the Windows page fault–handling code attempts to locate the page in the system paging file.

If the page can be located, it is loaded into physical memory, and the PTE is updated to reflect its location.

If it cannot be located and the translation is a user mode translation, an access violation occurs because the virtual address references an invalid physical address. If the page cannot be located and the translation is occurring in kernel mode, a bug check(also called a blue screen) occurs.

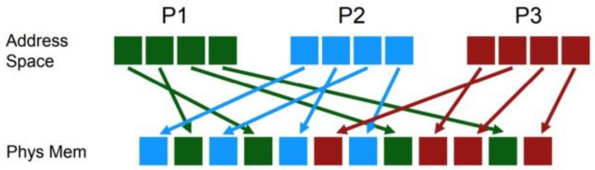


Figure 5. mapping relation

VIII. CONCLUSION

According to the characteristics of the tree, the maximum depth of balanced binary tree with n nodes is log2N, and the average search length is approximately log2N.

In this paper, our team discuss the software TLB management method based on balanced binary search tree through the abstract model. This management method takes advantage of the advantages of balanced binary search tree in pruning and searching, which can quickly complete the retrieval of virtual page number and physical page .

When the TLB expands, the number of TLB failures decreases, and the cost of tree building is shared equally in each search. To a certain extent, it can improve the performance of computer operating system in processing large virtual memory space, accelerate the conversion from virtual address to physical address, and make room for other designs on chip.

However, due to the lack of experimental environment and sufficient theoretical knowledge, our team only verify theoretical time compleity by math and statistic, did not test in the actual environment, especially not verify the program model ,which is based on balanced binary search tree, our team did not test reliability under the condition of hardware instability or storage space disorder, like no electrical power and other special circumstances.

A reliable software model which can be used in engineering must consider all kinds of special circumstances. Stability and reliability is the most important aspect of a software model. Balanced binary search tree as the search scheme of this model has the possibility of further optimization.

The algorithm designed in the future research should not only consider the running speed, but also consider the difficulty of implementation in practical engineering and the future maintenance work. It is necessary to do further study of complexity of maintenance work and the stability of the software model.

Engineering and theory are the unity of opposites. Theory is the source of engineering. The development of theory needs the support of engineering.

REFERENCES

[1] Marin G, Mellorcrummey J. Cross-architecture performance predictions for scientific applications using parameterized models[C]. measurement and modeling of computer systems, 2004, 32(1): 2-13.

[2] Boncz P, Manegold S, Kersten M L, et al. Database Architecture Optimized for the New Bottleneck: Memory Access[C]. very large data bases, 1999: 54-65.

[3] G.M.Adelson-Velsky,E.M.Landis, “An algorithm for the organization of information” 1962 30(1): 7-13

[4] Deb, Kalyanmoy, and Ram Bhushan Agrawal. "Simulated Binary Crossover for Continuous Search Space.." Complex Systems 1995.

[5] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[6] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] Talluri M, Hill M D. Surpassing the TLB performance of superpages with less operating system support[C]. architectural support for programming languages and operating systems, 1994, 29(11): 171-182.

[8] Sleator, Daniel D., and Robert E. Tarjan. "Self-adjusting binary search trees." Journal of the ACM 32.3 1985

[9] Rajwar R, Herlihy M, Lai K K, et al. Virtualizing Transactional Memory[C]. international symposium on computer architecture, 2005, 33(2): 494-505.

[10] Menon A, Santos J R, Turner Y, et al. Diagnosing performance overheads in the xen virtual machine environment[C]. virtual execution environments, 2005: 13-23.

Data Visualization Analysis of COVID-19 Epidemic Situation

Yang Tao
Nanchang Institute of Science and Technology
Nanchang, China
E-mail: taoyangxp@163.com

***Abstract*—2019 Novel Coronavirus (COVID-19) has brought immeasurable losses and huge impact to the world. For human health, many Centres for Disease Control(CDC) in various countries around the world are actively collecting data and doing a good job in virus prevention and control. The real-time release of the epidemic situation, with analysis and prediction, is a very effective method to combat the epidemic. By studying the situation of epidemic data, based on Jupyter Notebook, this paper gives the visual analysis process of COVID-19 epidemic data, and carries out specific analysis and implementation. And then it estimates the coronavirus converges roughly using sigmoid fitting. Although the sigmoid fitting tend to underestimate the curve, its actual value tend to be more than sigmoid curve estimation. The proposed data visualization analysis method could effectively display the status of the COVID-19 epidemic situation, hoping to help control and reduce the impact of the COVID-19 epidemic.**

Keywords—COVID-19; Data Visualization; Situation Dashboard; Jupyter Notebook; Epidemic Situation

I. INTRODUCTION

2019 Novel Coronavirus (2019-nCoV) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness. A growing number of patients reportedly have indicated person-to-person spread is occurring. At this time, it’s unclear how easily or sustainably this virus is spreading between people.

Therefore, it is very important to visually analyse the COVID-19 Epidemic Situation, which helps to control the impact of the COVID-19 epidemic and reduce losses.

II. LOAD DATA

The dataset has daily level information on the number of affected cases, deaths and recovery from 2019 novel coronavirus. This is a time series data and so the number of cases on any given day is the cumulative number. The data is available from 22 Jan, 2020. We can download latest data from Johns Hopkins University github repository: https://github.com/CSSEGISandData/COVID-19[1].We can also grab data from various Centres for Disease Control [2-6].

The data folder contains the previously posted dashboard case reports from Jan 21 to Feb 14, 2020 for the coronavirus COVID-19 (formerly known as 2019-nCoV). We will refer to the data provided in the new folder, entitled "csse_covid_19_data folder". Moving forward they will be updating daily case reports into this new folder. Additionally, the previously uploaded data from Jan 21-Feb 14, 2020 is also included in the new folder, and it has been cleaned and re-formatted to address inconsistencies in the time zone and update frequency that resulted during the transition from our manual updates to automated updates (which took place on Feb 1, 2020. The new folder now includes one case report per day, from the same time of day. This will be the standard moving forward (as of Feb 14, 2020). That is the data we will load for visualization analysis.

Main file in this dataset is covid_19_data.csv and the detailed descriptions are below.

- ✧ Sno - Serial number
- ✧ ObservationDate - Date of the observation in MM/DD/YYYY. We will convert ObservationDate and Last Update to datetime since they are currently taken as object.

DOI: 10.21307/ijanmc-2020-038

- Province/State - Province or state of the observation (Could be empty when missing)
- Country/Region - Country of observation
- Last Update - Time in UTC at which the row is updated for the given province or country. (Not standardised and so please clean before using it)
- Confirmed - Cumulative number of confirmed cases till that date
- Deaths - Cumulative number of deaths till that date
- Recovered - Cumulative number of recovered cases till that date

III. VISUALIZATION ANALYSIS

For the purpose of data visualization, we mainly use the Python-based tools of Jupyter Notebook[7] and plotly[8]. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. The plotly visualization is heavy used in this kernel so that we can interactively see the figure, map etc. As a side effect, it might take a little bit more time to initialize the Python environment and to load the kernel. Then grab data from the Internet and load the data.

A. Worldwide Trend

When we see the confirmed cases in worldwide, it just look like exponential growth curve. The number is increasing very rapidly especially recently. As a further matter, daily new confirmed cases started not increasing from April 4. After that, flat trend continues so far, as shown in Figure 1.

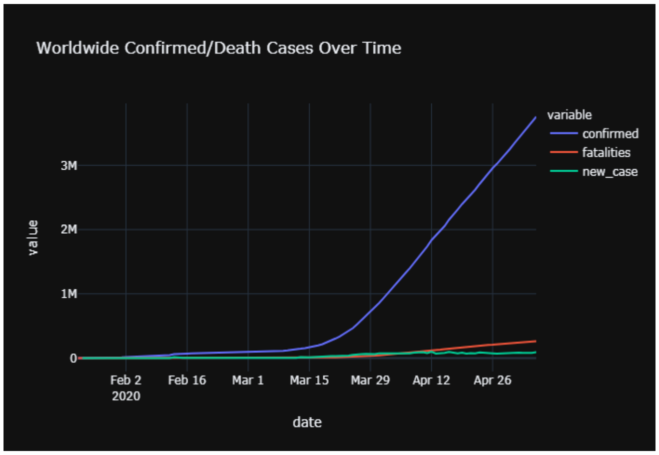


Figure 1. Worldwide Confirmed/Death Cases Over Time

Moreover, when we check the growth in log-scale below figure, we can see that the speed of confirmed cases growth rate slightly increases when compared with the beginning of March and end of March. In spite of the Lockdown policy in Europe or US, the number is still increasing rapidly, as shown in Figure2.

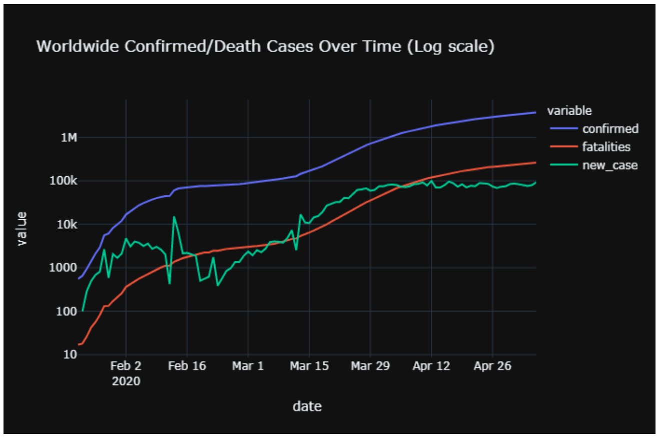


Figure 2. Worldwide Confirmed/Death Cases Over Time (Log scale)

It looks like fatalities curve is just shifted the confirmed curve to below in log-scale, which means mortality rate is almost constant. We see that mortality rate is kept almost 3%, however it is slightly increasing gradually to go over 7% at the end of April. Europe & US has more seriously infected by Coronavirus recently, and mortality rate is high in these regions, as shown in Figure 3. It might be because when too many people get coronavirus, the country cannot provide enough medical treatment.

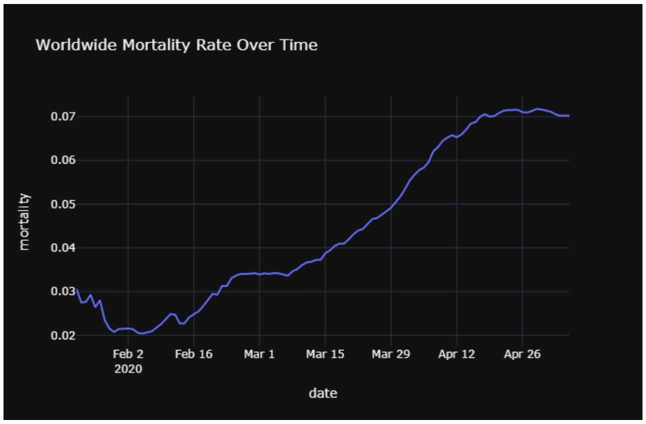


Figure 3. Worldwide Mortality Rate Over Time

B. Country-wise Growth

There are 187 countries in the dataset. How's the distribution of number of confirmed cases by country? It is difficult to see all countries so let's check top countries as shown in Figure 4.

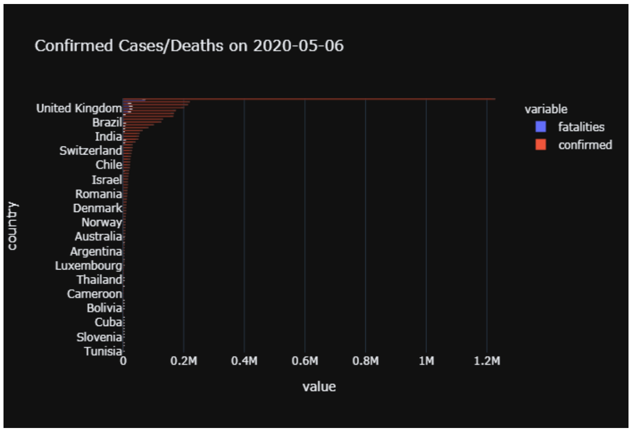


Figure 4. Confirmed Cases/Deaths on 2020-05-06

Now US, Italy and Spain has more confirmed cases than China, and we can see many Europe countries in the top. Korea also appears in relatively top despite of its population, this is because Korea executes inspection check aggressively.

Let's check these major country's growth by date.

As we can see, Coronavirus hit China at first but its trend is slowing down in March which is good news. Bad news is 2nd wave comes to Europe (Italy, Spain, Germany, France, UK) at March. But more sadly 3rd wave now comes to US, whose growth rate is much faster than China, or even Europe. Its main spread starts from middle of March and its speed is faster than Italy. Now US seems to be in the most serious situation in terms of both total number and spread speed. Now let's see the confirmed cases for the top 30 countries, as shown in Figure 5.

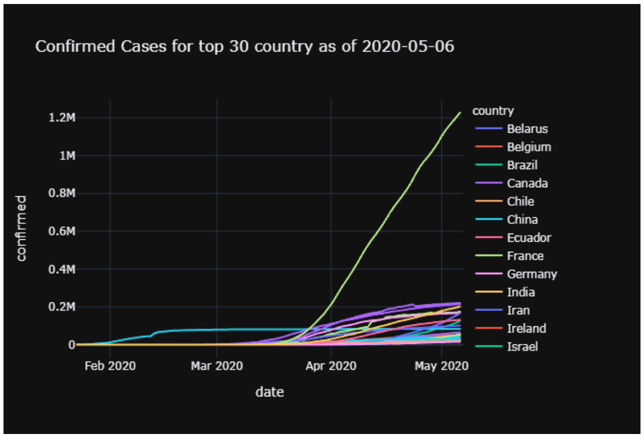


Figure 5. Confirmed Cases for top 30 country as of 2020-05-06

In terms of number of fatalities, Europe & US are serious situation now, as shown in Figure 6. Many countries have more fatalities than China now, including US, Italy, Spain, France, UK, Iran Belgium, Germany, Brazil, Netherlands. US's spread speed is the fastest, US's fatality cases become top1 on Apr 10th.

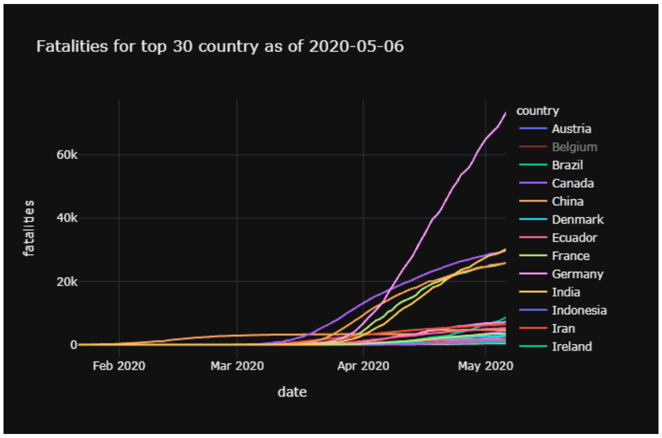


Figure 6. Fatalities for top 30 country as of 2020-05-06

Now let's see mortality rate by country, as shown in Figure 7.

Italy is the most serious situation, whose mortality rate is over 10% as of 2020/3/28.We can also find countries from all over the world when we see top mortality rate countries, as shown in Figure 7. Iran/Iraq from Middle East, Philippines & Indonesia from tropical areas. Spain, Netherlands, France, and UK form Europe etc. It shows this coronavirus is really worldwide pandemic.

The countries whose mortality rate is low are shown in Figure 8.

By investigating the difference between above & below countries, we might be able to figure out what is the cause which leads death.

Be careful that there may be a case that these country's mortality rate is low due to these country does not report/measure fatality cases properly.

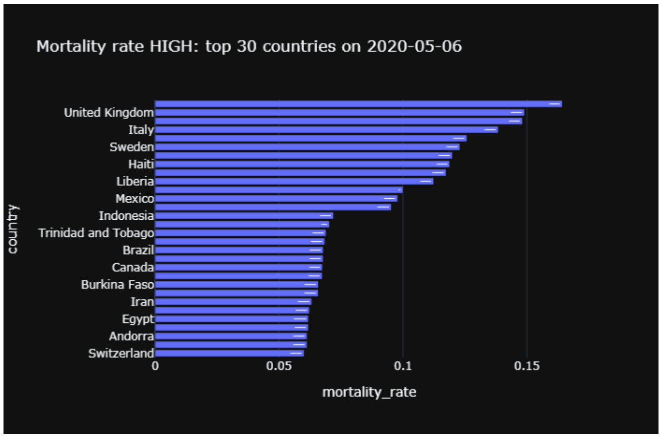


Figure 7. Mortality rate HIGH: top 30 countries on 2020-05-06

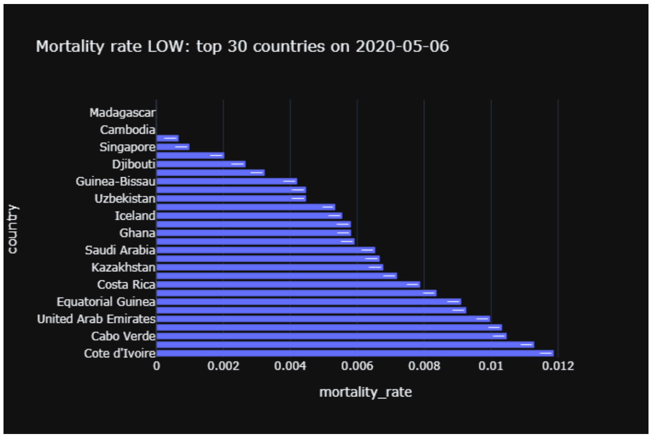


Figure 8. Mortality rate HIGH: top 30 countries on 2020-05-06

Let's see number of confirmed cases on map. Again we can see Europe, US, Middle East (Turkey, Iran) and Asia (China, Korea) are red, as shown in Figure 9.

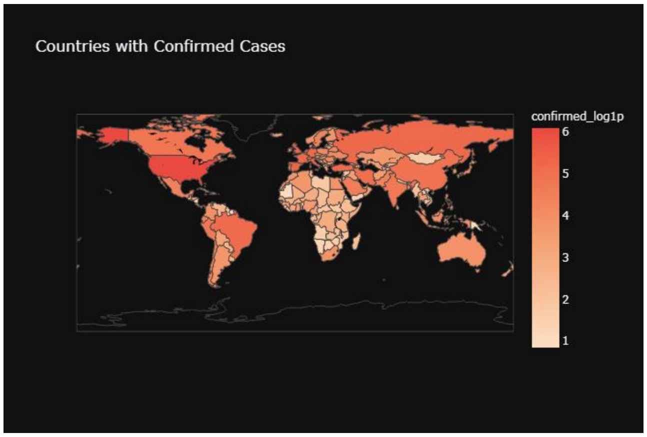


Figure 9. Countries with Confirmed Cases on 2020-05-06

The number of fatalities on map is shown as Figure 10 and the mortality rate map is shown as Figure 11.

When we see mortality rate on map, we see Europe (especially Italy) is high. Also we notice Middle East (Iran, Iraq) is high. When we see tropical area, I wonder why Philippines and Indonesia are high while other countries (Malaysia, Thai, Vietnam, as well as Australia) are low. For Asian region, Korea's mortality rate is lower than China or Japan, I guess this is due to the fact that number of inspection is quite many in Korea[9-10].

From the mortality rate map, it seems that mortality rate is especially high in Europe region, compared to US or Asia.

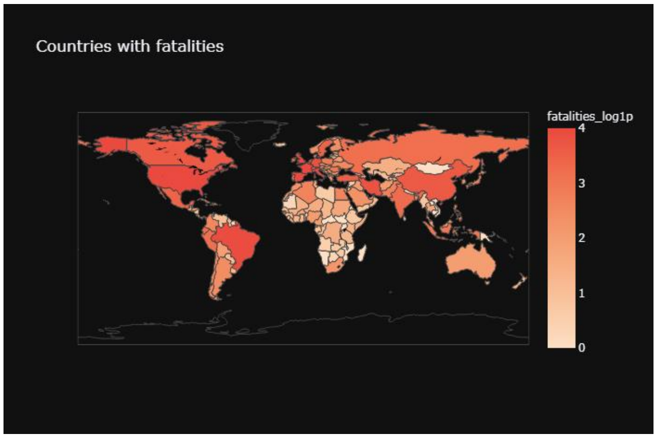


Figure 10. Countries with fatalities on 2020-05-06

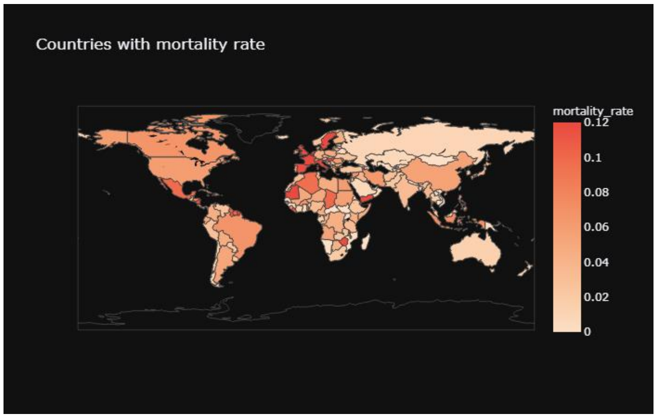


Figure 11. Countries with mortality rate on 2020-05-06

Why mortality rate is different among country? What kind of hint is hidden in this map? Especially mortality rate is high in Europe and US, is there some reasons? There is one interesting hypothesis that BCG vaccination[11].

C. Daily NEW Confirmed Cases Trend

Let's see the DAILY new cases trend as shown in Figure12.

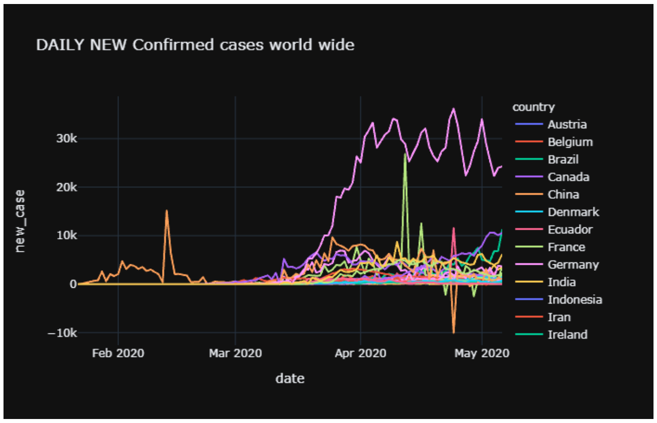


Figure 12. DAILY NEW Confirmed cases worldwide

We find from the figure 12:

- China has finished its peak at Feb 14, new confirmed cases are surpressed now.
- Europe&US spread starts on mid of March, after China slows down.
- As effect of lock down policy in Europe (Italy, Spain, Germany, France) now comes on the figure, the number of new cases are not so increasing rapidly at the end of March.
- Current US new confirmed cases are the worst speed, recording worst speed at more than 30k people/day at peak. Daily new confirmed cases start to decrease from April 4 or April 10.
- After that we can see a weekly trend that the confirmed cases becomes small on Monday. I think this is because people don't (or cannot) get medical care on Sunday so its reporting number is low on Sunday or Monday.

D. Zoom up to US

As we can see, the spread is fastest in US now, at the end of March. Let's see in detail what is going on in US. When we see inside of the US, we can see only New York, and its neighbour New Jersey dominates its spread and are in serious situation. The number of New York confirmed cases is over 50k, while other states are less than about 5k confirmed cases, as shown in Figure 13.

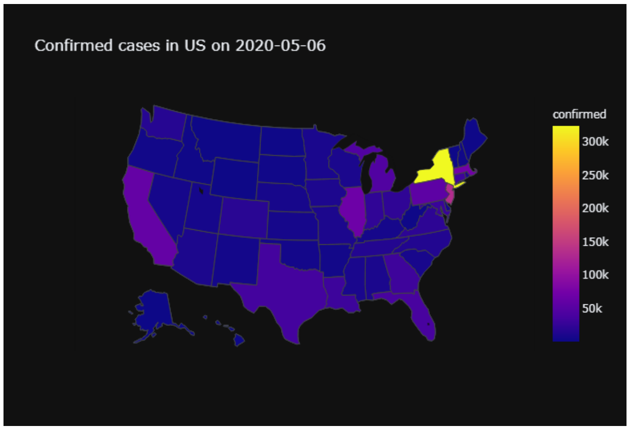


Figure 13. Confirmed cases in US on 2020-05-06

Mortality rate in New York seems not high, around 2% for now, as shown in Figure 14.

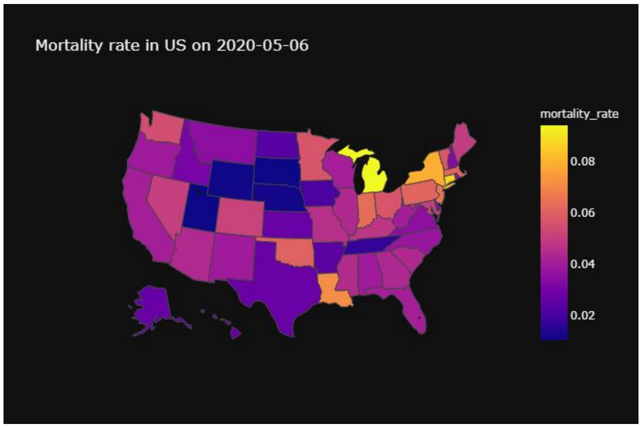


Figure 14. Mortality rate in US on 2020-05-06

All state is US got affected from middle of March, and now growing exponentially. In New York, less than 1k people are confirmed on March 16, but more than 50k people are confirmed on March 30. 50 times explosion in 2 weeks! The confirmed cases by state in US is show in Figure 15.

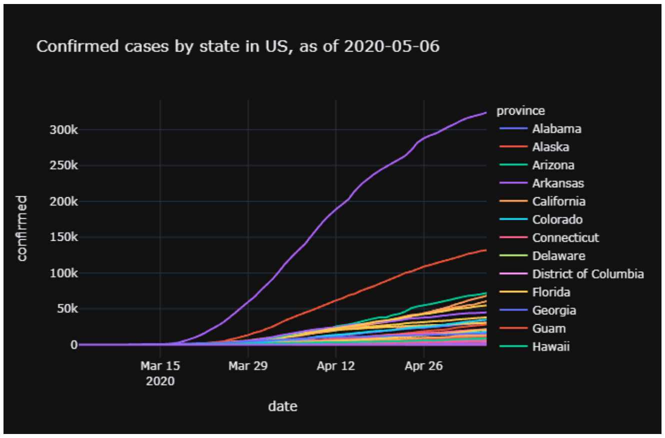


Figure 15. Confirmed cases by state in US, as of 2020-05-06

E. Zoom up to Europe

When we look into the Europe, its Northern & Eastern areas are relatively better situation compared to Eastern & Southern areas. The map of European Countries with Confirmed Cases is shown as Figure 16 and Figure 17.

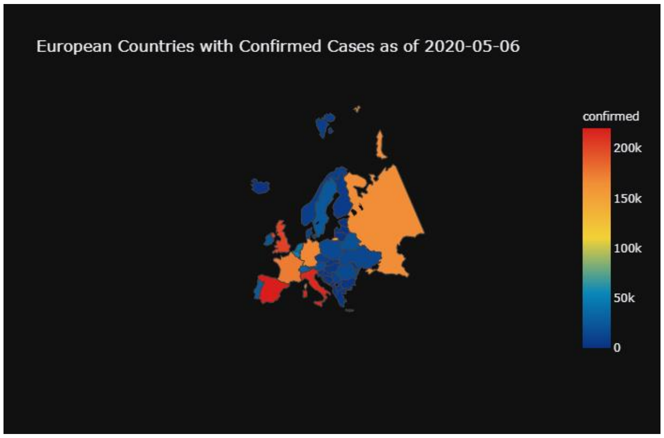


Figure 16. European Countries with Confirmed Cases, as of 2020-05-06

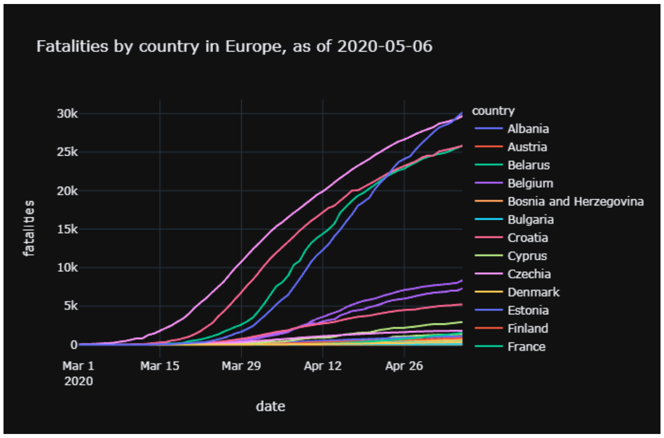


Figure 17. Confirmed cases by country in Europe, as of 2020-05-06

Especially Italy, Spain, German, France, UK are in more serious situation. Number of confirmed cases rapidly increasing in Russia now (as of May 1), Russia is now potentially very dangerous situation.

When we check daily new cases in Europe(as shown in Figure 18), we notice:

- UK and Russia daily growth are more than Italy now. These countries are potentially more dangerous now.
- Italy new cases are not increasing since March 21, maybe due to lock-down policy is started working.

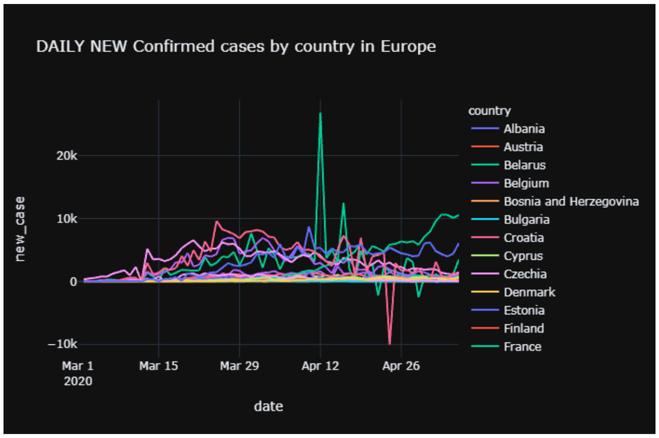


Figure 18. DAILY NEW Confirmed cases by country in Europe

F. Zoom up to Asia

In Asia, China & Iran have many confirmed cases, followed by South Korea & Turkey. Asian Countries with Confirmed Cases is as shown in Figure 19.

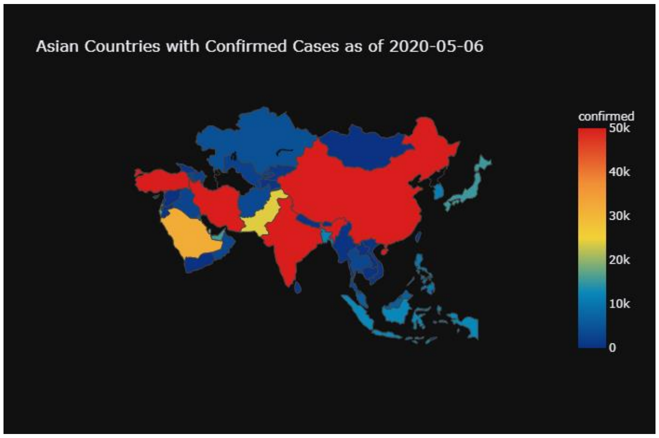


Figure 19. Confirmed cases by country in Asia, as of 2020-05-06

The coronavirus hit Asia in early phase, how is the situation now?

China & Korea is already in decreasing phase. Unlike China or Korea, daily new confirmed cases were kept increasing on March or April, especially in Iran or Japan. But the number is started to decrease now on these country as well, as shown in Figure 20.

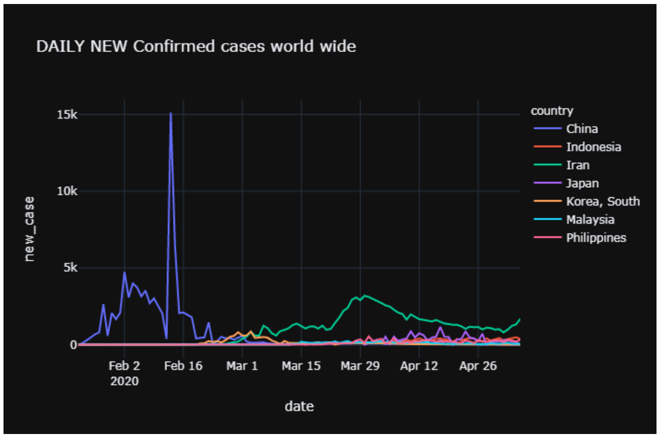


Figure 20. DAILY NEW Confirmed Cases in Asia, as of 2020-05-06

IV. ESTIMATION

Of course everyone is wondering when the coronavirus converges. Let's estimate it roughly using sigmoid fitting.

I referenced two kernels[12-13] for original ideas. The fitting result is shown in Figure 21.

Sigmoid fitting with all latest data

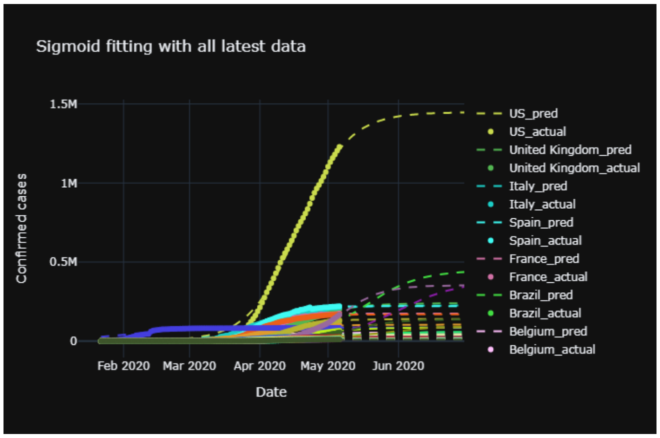


Figure 21. Sigmoid fitting with all latest data

If believe above curve, the number of confirmed cases is slowing down now and it will be converging around the beginning of May in most of the country. It might take until beginning on June in US.

Let's try validation by excluding last 7 days data, as shown in Figure 22.

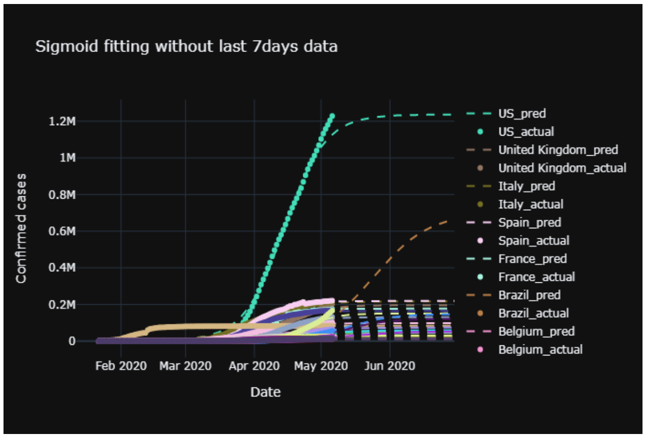


Figure 22. Sigmoid fitting without last 7days data

Now noticed that sigmoid fitting tend to underestimate the curve, and its actual value tend to be more than sigmoid curve estimation.

Therefore, need to be careful to see sigmoid curve fitting data; actual situation is likely to be worse than the previous figure trained with all data.

V. CONCLUSION

Based on data available on May 6, the paper showed the visualization of the COVID-19 Epidemic Situation, including the worldwide trend, country-wide growth, and so on. Then it estimated when the coronavirus converges roughly using sigmoid fitting. The model's estimates and predictions closely match reported confirmed cases. Therefore the proposed data visualization analysis method could effectively display the status of the COVID-19 epidemic situation, hoping to help control and reduce the impact of the COVID-19 epidemic.

The next steps include applying the method to global COVID-19 death data into small regions, as provinces. The method of visualization analysis could also be used to evaluate population mortality and the spread of other diseases.

ACKNOWLEDGMENTS

The authors wish to thank Corochann, who is an Engineer at Preferred Networks in Tokyo. This work was supported in part by the Ph.D. Research Initiation Fund of Nanchang Institute of Science and Technology with the Project (No. NGRCZX-18-01) and supported by Robot and Intelligent System Research Centre of Artificial Intelligence College of Nanchang Institute of Science and Technology (No.NGYZY-20-005).

REFERENCES

[1] https://www.kaggle.com/benhamner/covid-19-forecasting-challenges-week-2-data-prep.

[2] China CDC (CCDC): http://weekly.chinacdc.cn/news/TrackingtheEpidemic.htm

[3] Taiwan CDC: https://sites.google.com/cdc.gov.tw/2019ncov/taiwan?authuser=0

[4] US CDC: https://www.cdc.gov/coronavirus/2019-ncov/index.html

[5] Government of Canada: https://www.canada.ca/en/public-health/services/diseases/coronavirus.html

[6] European Centre for Disease Prevention and Control (ECDC): https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

[7] Jupyter notebook - Project Jupyter | Home. https://jupyter.org/

[8] Plotly: The front-end for ML and data science models. https://plotly.com/

[9] South Korea launches 'drive-thru' coronavirus testing facilities as demand soars. https://www.japantimes.co.jp/news/2020/03/01/asia-pacific/science-health-asia-pacific/south-korea-drive-thru-coronavirus/#. XoAmw4j7RPY

[10] Coronavirus: Why Japan tested so few people. https://asia.nikkei.com/Spotlight/Coronavirus/Coronavirus-Why-Japan-tested-so-few-people

[11] If I were North American/West European/Australian, I will take BCG vaccination now against the novel coronavirus pandemic. https://www.jsatonotes.com/2020/03/if-i-were-north-americaneuropeanaustral.html

[12] Sigmoid per country. https://www.kaggle.com/group16/sigmoid-per-country-no-leakage

[13] COVID-19 growth rates per country. https://www.kaggle.com/mikestubna/covid-19-growth-rates-per-country.

Conditional GAN-based Remote Sensing Target Image Generation Method

Liu Haoyang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, ShaanXi, China
E-mail: 502339341@qq.com

Hu Zhiyi
Engineering Design Institute
Army Research Loboratory
Beijing, 100000, China
E-mail: 763757335@qq.com

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

Gao Shouyi
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, ShaanXi, China
E-mail: 478204287@qq.com

***Abstract*—In view of the uncontrollable process of traditional GAN generating remote sensing target images, the shortcomings of generated samples are similar, and lack diversity. This paper proposes a generative confrontation network model based on background conditions. First, the computer vision attention mechanism is introduced into the generative confrontation network. Choose a learning target model so that the GAN network only learns the target model during training, and ignores other non-target information. Reduce the dependence on the number of samples in the GAN training process; secondly, use the U-net network as a generator to restore other non-target information when generating the remote sensing image of the target as much as possible; again, distinguish by different colors of the conditional mask The category of the generated target; at the same time, the L1 regularization loss is added to the loss term of the generator model, and finally, the remote sensing target image is generated. The experimental results show that the peak signal-to-noise ratio (PSNR) of the remote sensing image generation algorithm proposed in this paper reached 18.512, and the structural similarity (SSIM) reached 88.47%, which is better than the comparison test model where the generator is an ordinary autoencoder.**

Keywords-Remote Sensing Target Image; Conditional GAN; U-net; Conditional Mask

I. INTRODUCTION

In recent years, due to the advancement and maturity of satellite remote sensing technology, target detection based on deep neural network methods [1-2] has played a more significant role in the search and rescue of sea ships and the detection and recognition of sea ship types [3-6]. However, an accurate and generalized remote sensing target detection algorithm needs to be trained on the largest possible data set. However, in reality, the cost of acquiring a large number of remote sensing images is extremely high, so it is necessary to use data enhancement technology to expand the existing image samples.

The image in the original sample undergoes rotation, translation, cropping, etc., to generate new image training data. Although this method increases the number of training samples, the samples are highly similar, without increasing the diversity of the sample set. Although the generator model of the Variational Autoencoder (VAE) [7-9] can increase the diversity of the data set, the generated image is blurred and the effect is not good in the process of training the network, so it is not suitable for increasing the remote sensing image data Set method.

Ian J. Goodfellow et al. proposed Generative adversarial nets (GANs) [10]. The powerful fitting ability is provided by the multilayer neural network. In

DOI: 10.21307/ijanmc-2020-039

theory, it is possible to model any data distribution and build a complex data model [11]. But this kind of network needs more target samples when generating remote sensing target images, and the generation process is uncontrollable. It is easy to cause problems such as high similarity of generated samples and lack of diversity in samples.

To optimize the above problems, this paper proposes a conditional generative adversarial network [15] (Conditional Generative Adversarial Nets, CGAN for short). Here, this article takes the ship target on the sea as an example. When there are only a few remote sensing image samples of the ship, the generated ship samples are controlled by the ocean background conditions. This method not only increases the number of samples of remote sensing images of sea ships, but also improves the imaging quality of ship samples, thereby effectively expanding the data set of remote sensing target images, improving the generalization capabilities of classification and detection models, and making them more Good application in the field of target detection.

II. CONDITIONAL GENERATIVE COUNTERMEASURE NETWORK MODEL

A. *The basic principle of conditional generative countermeasure network*

The traditional generative confrontation network has the characteristics of directly obtaining data distribution. "Real data" can be generated without pre-modeling the data distribution in advance, but the process of its generation is uncontrollable, and the generated samples are often similar and lack diversity. To make the generated samples efficient and controllable, experts and scholars have made improvements based on traditional GAN. As shown in Figure 1, both the generator and the discriminator add a constraint y, which can be any meaningful information. Input y as an additional input into the generator together with the prior noise z, which affects the generated data. The discriminator will also give a prediction under the influence of the constraint y, to achieve control of the generated samples. The objective optimization function definition of the conditional generative confrontation network is shown in formula 1:

$$\min_{G}\max_{D} V(G,D) = E_{x\sim p_{data(x)}}\left[\log D(x|y)\right] + E_{z\sim p_z}\left[\log(1 - D(G(x|y)))\right] \quad (1)$$

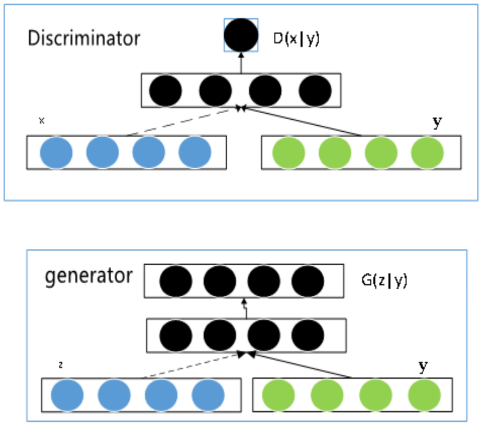


Figure 1. Conditional generative countermeasure network structure

B. *Model design of conditional generative countermeasure network*

Due to the high cost of acquiring remotely sensed sea target image samples, the traditional DGAN model is not ideal when applied to sea target image generation. To reduce the dependence on the number of samples, this paper proposes a conditional generative confrontation network model for ocean background. This model introduces a visual attention mechanism. Use the mask to synthesize the conditional mask image with the ordinary sea surface image, and then control the generated data by using the conditional image as a constraint condition. For example, use the color value of the mask to control the type of ship-generated. The network only needs to learn the data distribution of the ship model during the training process, while ignoring the non-target sea background information, so that it can generate ship images that meet the conditional category with a small number of sample data sets. The network encodes the mask in the conditional image during training. A kind of mapping is formed in the hidden space, and then the corresponding target is generated by decoding. The principle of the network model in this paper is shown in Figure 2.

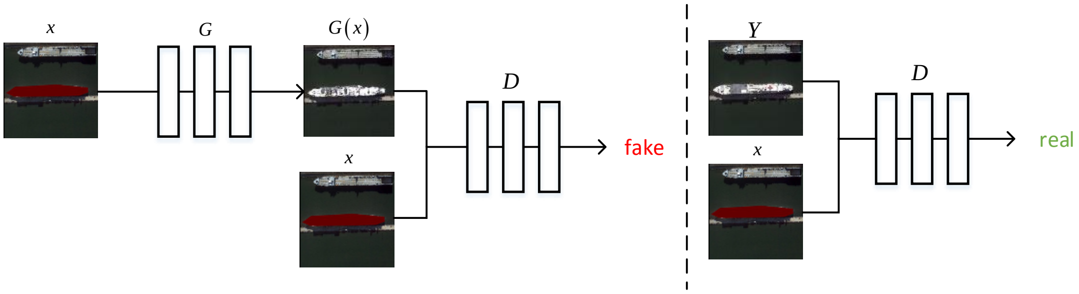


Figure 2. Working principle of marine background condition generated countermeasure network model

In order to realize the attention mechanism of task encoding→decoding, the generator of this paper chooses the network model of the U-net structure. The discriminator uses the structure of an area discriminator.

1) Network structure of generator model based on U-NET network

In order to obtain valuable information from the conditional background picture, the network structure of the generator in this paper adopts the U-net network structure of cascaded encoding and decoding. The U-net network was first proposed by Olaf et al. to be applied to the full convolutional network (FCN) of medical image segmentation. The U-Net generative model is similar to the traditional self-encoding model, including encoding and decoding structures. At the same time, a skip connection method is used between the i-th layer and n-i. The feature map of the i-th layer is connected with the output of the n-i-1th layer as the input of the n-ith layer. During output, the underlying image information that needs to be preserved in the background is restored, and the part of the ship mask that needs to be generated is reconstructed. The network structure of the U-Net generation model is shown in Figure 3.

Figure 3 shows that for the 256×256 size high-resolution remote sensing sea surface ship images in this article, the coding structure is an 8-layer down-sampling convolution operation. Each time the sample is down, the length and width of the feature map are reduced by half. The number of channels of the image sequentially becomes {64, 128, 256, 512, 512, 512, 512, 512}. The size of the input image is fixed at 256×256×3. After 8 layers of downsampling, the output size is 1×1×512. The decoding structure is an 8-layer up-sampling transposed convolution operation. Each up-sampling is connected to the output of the corresponding coding layer, the length and width of the feature map are doubled, and the number of image channels becomes {512, 1024, 1024, 1024, 1024, 512, 256, 128, 3}. Finally, a target generated image with a size of 256×256×3 is obtained.

The Unet network structure is divided into two parts: encoding and decoding. The coding network uses convolutional layers for down-sampling and extracts the semantic information in the conditional image. The decoding network uses transposed convolutional layers. Input a conditional background image into the encoding network. After extracting the image semantic information feature, it is output through the decoder. Generate qualified remote sensing images of ships. The specific parameters of the network are shown in Table 1.

2) Network structure of discriminator model

The discriminator network as a whole is still a traditional deep convolutional neural network. Different from the general discriminator, the output of the network is not a single prediction value of the true or false of the image, but a prediction matrix. But the discriminator is still a two-class network. For the output prediction matrix, the error from the label is calculated by the mean value, and the error is averaged and passed to the network as a loss. The network then updates the parameters according to the loss. Specifically, each item in the prediction matrix represents an N×N image block in the input. The discriminator finally gives the final prediction result of the input image by averaging the prediction results of all image blocks.

First, the input of the discriminator is to cascade the conditional image and the discriminant image to get an input tensor of 256×256×6. The network extracts features through the five-layer convolutional layer to obtain 30×30×1 output. When the identification network is trained, the input for the combination of the generated picture and the conditional mask is a negative sample, and the real picture and the conditional mask image are used as a positive sample.

In order for the network to generate good images, this article has made many attempts on the data set. It is finally determined that when the output size is 30×30, the effect of generating 256×256 remote sensing sea surface ship images is the best. The discriminator network structure is shown in Table 2 and Figure 4.

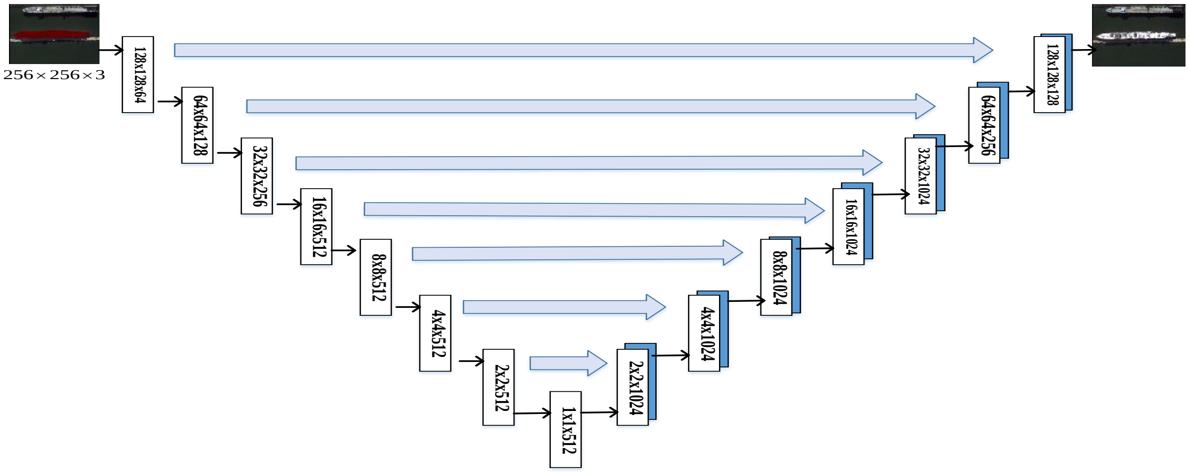


Figure 3. U-net generator network structure and coding and decoding structure diagram

TABLE I. GENERATOR MODEL NETWORK STRUCTURE PARAMETER TABLE

Inputs	Type	Kernel	Batch Normalization	Activation Function	Outputs
256x256	conv	4x4	YES	RELU	128x128
128x128	conv	4x4	YES	RELU	64x64
64x64	conv	4x4	YES	RELU	32x32
32x32	conv	4x4	YES	RELU	16x16
16x16	conv	4x4	YES	RELU	8x8
8x8	conv	4x4	YES	RELU	4x4
4x4	conv	4x4	YES	RELU	2x2
2x2	conv	4x4	YES	RELU	1x1
1x1	deconv	4x4	YES	RELU	2x2
2x2	deconv	4x4	YES	RELU	4x4
4x4	deconv	4x4	YES	RELU	8x8
8x8	deconv	4x4	YES	RELU	16x16
16x16	deconv	4x4	YES	RELU	32x32
32x32	deconv	4x4	YES	RELU	64x64
64x64	deconv	4x4	YES	RELU	128x128
128x128	deconv	4x4	YES	RELU	256x256

TABLE II. DISCRIMINATOR MODEL NETWORK STRUCTURE PARAMETER TABLE

Inputs	Type	Kernel	Batch Normalization	Activation Function	Outputs
256x256	conv	4x4	YES	LeakyReLU	128x128
128x128	conv	4x4	YES	LeakyReLU	64x64
64x64	conv	4x4	YES	LeakyReLU	32x32
32x32	conv	4x4	YES	LeakyReLU	31x31
31x31	conv	4x4	YES	LeakyReLU	30x30

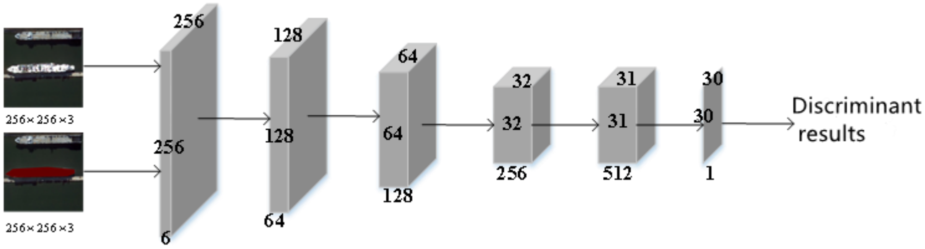


Figure 4. Network structure diagram of discriminator

3) Design of loss function

In this paper, the least square loss is used to improve the training instability problem in the generative confrontation network. For the generative model G and its corresponding discriminant model D, the loss function is defined as shown in formula 2.

$$L_{cGAN}(G,D)=\frac{1}{2}\left(E_{x\sim p_{data(y)}}[(D_Y(y)-1)^2]+E_{x\sim p_{data(x)}}[(D_Y(G(x|y)))^2]\right) \qquad (2)$$

In addition to the maximum and minimum optimization of the original loss of the generative confrontation network, the network also needs to generate the corresponding type of ship based on the mask in the conditional image, and also consider filling in the background image in the generated image. The sea background in the conditional image, so in order to minimize the difference between the generated image and the original background, the L1 regularization loss is added to the loss term of the generated model G. The loss function is defined as shown in formula 3.

$$L_{L1}(G)=E_{x,y,z}\left(\|y-G(x|y)\|_1\right) \qquad (3)$$

The final objective loss function becomes as shown in formula 4.

$$G^* = \arg\min_{G}\max_{D} L_{cGAN}(G,D)+\lambda L_{L1}(G) \qquad (4)$$

III. Verification Experiment and Comparative Analysis

In order to verify whether the conditional generation model in this paper can generate remote sensing target images in non-training samples, and to verify the quality of the generated images. We will generate random samples and compare and evaluate the sample images generated by the model in this article and the normal model.

A. Preparation of experimental environment

This experiment uses Linux operating system and python deep learning framework. The configuration of the specific experimental environment is shown in Table 3.

B. Design of experimental scheme

Figure 5 shows the flow of this experimental program. The experiment is mainly divided into the following main steps:

Firstly, the data source image is preprocessed; secondly, the trained model and the comparative experimental model are separately trained; thirdly, the randomly selected mask image is used as the input of the model to generate remote sensing target images; finally, the two models are generated Comparative evaluation of images. They are described separately below.

TABLE III. Experimental environment

Operating system	Ubuntu 18.04 LTS 64bit
CPU	Intel(R)Xeon(R) Gold 5118 CPU@2.30GHz
GPU	Nvidia GeForece TITAN Xp
Memory	32G
programing language	Python3.6.1
compiler	Pycharm2018.3
Deep learning framework	pytorch 0.4

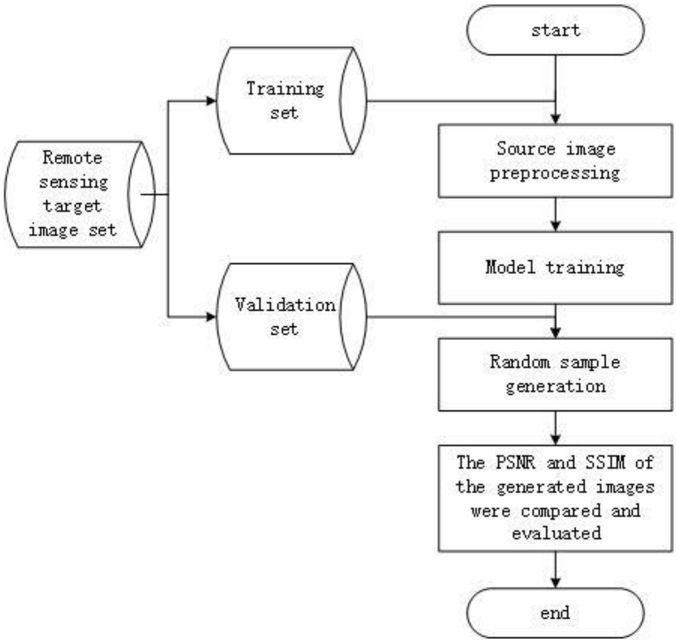

start
Training set
Remote sensing target image set
Validation set
Source image preprocessing
Model training
Random sample generation
The PSNR and SSIM of the generated images were compared and evaluated
end

Figure 5. Experimental process

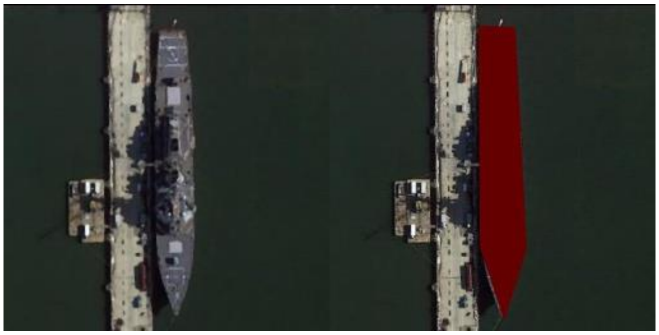
1) Preprocessing of data source image

In this paper, dozens of high-resolution remote sensing images of ships have been collected, and these images are relatively large. After appropriately cropping the image, resize it to 256*256. In order to be able to generate ship images in given background conditions. We masked the ships and aircraft carriers in the sample ship image by category, that is, the RGB values of the mask image pixels of the ship are (100, 0, 0) respectively. The pixel RGB values of the mask image of the aircraft carrier are respectively (0, 100, 0), and finally the mask image is paired with the target picture. The preprocessed 420 data sets are divided into training sets (336 sheets) and validation sets (84 sheets) at the ratio of 0.8 to 0.2. The preprocessed original image and conditional mask image are shown in Figure 6. Figure 6(a) is the real remote sensing target image, including sea background and ships, and Figure 6(b) is the conditional mask image corresponding to (a), in which the part of the ship is marked with a red mask.

2) Model training

It can be seen from Figure 7 that the generator learns the original image and the discriminator feedback information, and uses the mask image to generate the target image. The discriminator discriminates the image generated by the generator under the influence of the mask image until the discrimination prediction rate converges to 0.5.

In the model training process, different models use the same parameters when training on the remote sensing image data set. In order to make good use of each sample. The batch size during training is 1. Make the network fully learn the characteristics of ships in each sample; the initial learning rate is set to learning rate to 0.0002; the number of training iterations epoch is set to 1000; the Adam method is used as the training optimizer, and the momentum parameter is set to 0.5. The training time of the network proposed in this paper is about 5 hours.



（a） Real remote sensing target image （b）Conditional mask image

Figure 6. Preprocessing of high resolution ship remote sensing image

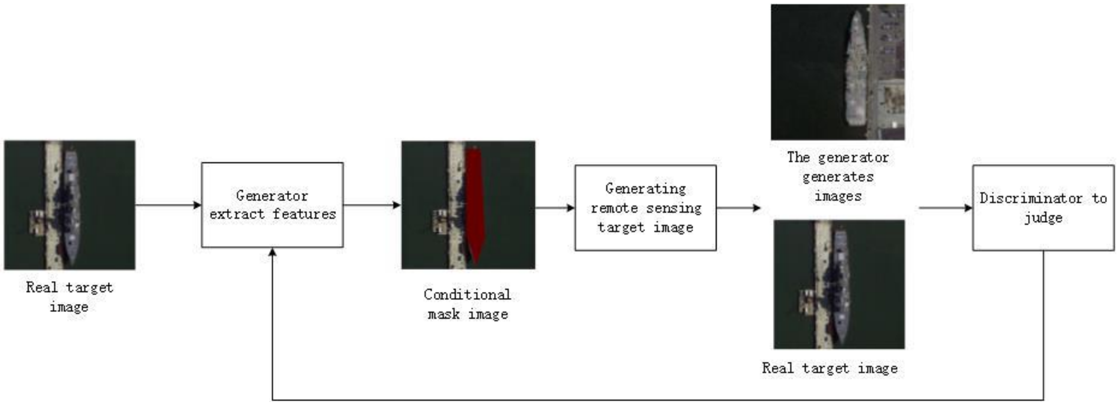
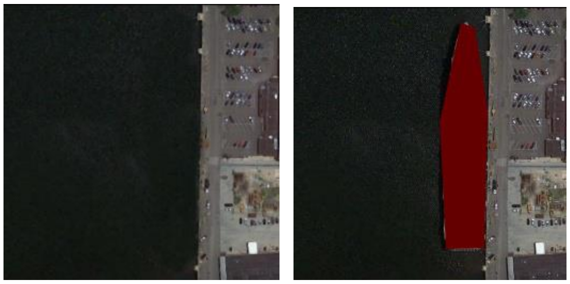

Real target image
Generator extract features
Conditional mask image
Generating remote sensing target image
The generator generates images
Real target image
Discriminator to judge

Figure 7. Model training process

3) *Random sample generation*

In order to test whether the generative confrontation network model proposed in this paper can generate remote sensing target images that are not in the training set. We will randomly select images from the images that are not involved in training and add any number of masks. Generate new remote sensing sea surface target images through the model in this paper. The masks of the ships and aircraft carriers used are derived from the set of ship masks saved when the training samples were made. An example of labeling is shown in Figure 8.



(a) Original image (b)Mask image

Figure 8. Sample random sample conditional mask

4) *Comparison and evaluation of generated images*

To detect the quality of the generated image, we need to compare and evaluate the sample image generated by the model in this article and the ordinary model. The evaluation is mainly to compare the peak signal-to-noise ratio and structural similarity. Randomly select images from the data validation set. After the remote sensing sea surface target image is generated, the peak signal-to-noise ratio and structural similarity value of the real image are calculated.

Compared with the experimental model, the U-net in the model is replaced with an ordinary autoencoder. The self-encoding network is divided into three parts: encoder, converter and decoder. The function of the encoder is to extract the feature and semantic information of the conditional background image. The converter maps the feature and semantic information extracted by the encoder to the hidden space and saves the feature information and semantic information. The decoder network will decode the activated features and semantic information from the converter network, generate pictures layer by layer, and get our final generated pictures.

The network structure of the self-encoding generator is shown in Figure 9.

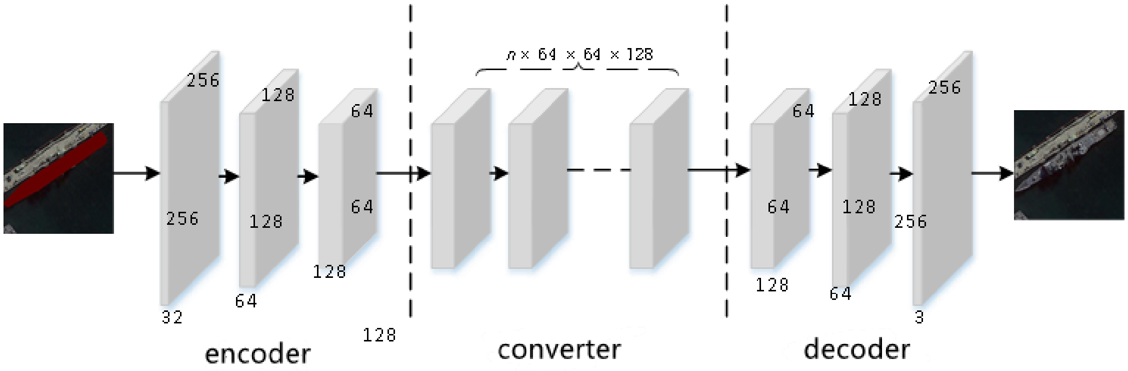


Figure 9. Generator structure of contrast experiment model

The encoder part also uses a deep convolutional network. Whenever down-sampling is performed through the convolutional layer, the scale of the network output is halved. The number of feature channels has been doubled. The decoder part also uses a deep transposed convolutional network. After transposing the convolutional layer once, the size of the image output by the network is doubled and the number of channels is halved. In classification and recognition tasks, Resnet (deep parameter transfer) and Densenet (dense) are conventional network design ideas. Both structures can deepen the depth of the network. Strengthen the transfer of features, make full use of features, and at the same time alleviate the problem of gradient disappearance due to depth during network training. Because of these advantages, the residual network structure used in the converter part is designed to contain 9 and 6 residual block converters respectively. The residual block network structure in the converter is shown in Figure 10.

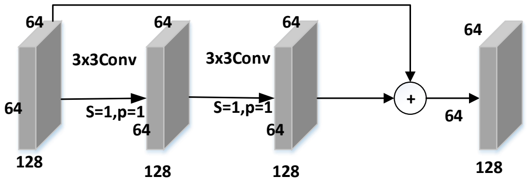
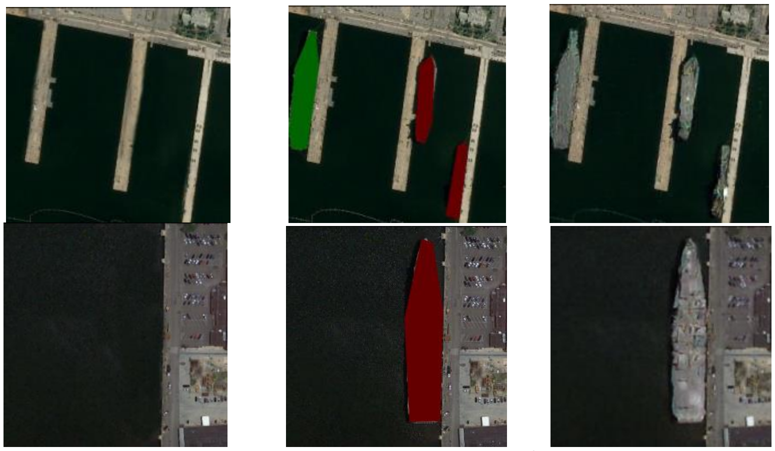


Figure 10. Residual block network structure in converter

The specific steps for the comparison and evaluation of the generated images are as follows: (1) 30 images are randomly selected from the 84 samples in the verification set, and the corresponding masks are marked. (2) The conditional mask image is used to generate remote sensing sea surface target images through the model and comparison model in this paper. (3) Calculate the peak signal-to-noise ratio and structural similarity of the images produced by different algorithms, and calculate the average value as the final evaluation.

C. *Experimental results and analysis*

Figure 11 shows a sample image of a ship of the corresponding category generated by the method in this paper through the given condition image. The generation algorithm proposed in this paper, after learning the previous samples, adds conditional masks under the new background, the model can generate new remote sensing sea ship images according to the mask type. It is proved that in the case of a few samples, more remote sensing sea surface target images can be generated by the method in this paper.



(a)Real picture (b)Conditional mask image (c)The results of the model generation in this paper

Figure 11. Random conditional sample generation results

In the comparative evaluation experiment, the peak signal-to-noise ratio and structural similarity of the images generated by this model and the comparative experimental model were calculated.

As shown in Table 4, the peak signal-to-noise ratio values of different generative model algorithms are compared. It can be seen from the table that the same training set and test set are used. Under the same experimental conditions, the image generation algorithm proposed in this paper is higher than the other two algorithms in terms of image authenticity, and the noise level is low.

TABLE IV. COMPARISON OF PSNR EVALUATION RESULTS

Image generation network	PSNR
The method of this paper	18.512
A generation model with nine residual blocks	17.596
A generating model with six residual blocks	17.089

As shown in Table 5, the structural similarity between the generated image of different model algorithms and the real image is compared. It can be seen from the table that training under the same training set, the structural similarity between the generated image and the real image in this article is better than other comparative experimental models.

TABLE V. COMPARISON OF PSNR EVALUATION RESULTS

Image generation network	SSIM
The method of this paper	88.47%
A generation model with nine residual blocks	81.65%
A generating model with six residual blocks	76.31%

IV. CONCLUSION

This paper proposes a conditional generative countermeasure network that can generate corresponding category high-resolution remote sensing ship image samples based on conditional semantic information in the background under small samples. And carried out the relevant experimental verification, at the same time, the objective and quantitative evaluation of this algorithm based on the conditional background image generation. It briefly introduces the relevant basic theory of the evaluation method. The image generation algorithms of different generator structures are compared through experiments. Under the same conditions of the background image, the peak signal-to-noise ratio (PSNR) of the image generation algorithm proposed in this paper reached 18.512, and the structural similarity (SSIM) reached 88.47%, which is better than the general autoencoder. The comparative test model.

REFERENCES

[1] Tan Kun, Wang Xue, Du Peijun. Remote sensing image classification based on deep learning and semi supervised learning [J]. Chinese Journal of image graphics, 2019, 24 (11): 1823-1841Zhang Zemiao, Huo Huan, Zhao Fengyu. Review of Object detection algorithms for deep convolutional neural networks [J]. Minicomputers, 2019, 40(09): 1825-1831.

[2] Liu Xiaobo, Liu Peng, Cai Zhihua, Qiao Yulin, Wang Ling, Wang min. research progress of optical remote sensing image target detection based on deep learning [J / OL]. Acta automatica Sinica: 1-13 [2019-12-23] https://doi.org/10.16383/j.aas.c190455.

[3] Chen Huiyuan, Liu Zeyu, Guo Weiwei, Zhang Zenghui, Yu Wenxian. Fast detection method of ship target in large scene remote sensing image based on cascaded convolutional neural network [J]. Acta radari Sinica, 2019, 8 (03): 413-424

[4] Yin ya, Huang Hai, Zhang Zhixiang. Research on ship target detection technology based on optical remote sensing image [J]. Computer science, 2019, 46 (03): 82-87

[5] Xu Fang. Research on key technologies for automatic detection of sea surface targets in visible remote sensing images [D]. University of Chinese Academy of Sciences (Changchun Institute of optics, precision machinery and physics, Chinese Academy of Sciences), 2018

[6] Yang Miao. Research on ship detection algorithm in port based on optical remote sensing image [D]. Xihua University, 2018

[7] Kingma, Diederik P, Welling, Max. Auto-Encoding Variational Bayes[J].

[8] Kingma, Diederik P, Rezende, Danilo J, Mohamed, Shakir, and so on. Semi-Supervised Learning with Deep Generative Models[J]. Advances in Neural Information Processing Systems, 2014, 4:3581-3589.

[9] Bodin, Erik, Malik, Iman, Ek, Carl Henrik, and so on. Nonparametric Inference for Auto-Encoding Variational Bayes[J].

[10] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[11] Gan LAN, Shen Hongfei, Wang Yao, Zhang Yuejin. Data set enhancement method based on improved dcgan [J / OL]. Computer application: 1-11 [2020-11-10] http://kns.cnki.net/kcms/detail/51.1307.TP.20201015.1715.017.html.

[12] Lin Zhipeng, Zeng Libo, Wu qiongshui. Cervical cell image data enhancement based on generative antagonism network [J]. Science, technology and engineering, 2020,20 (28): 11672-11677

[13] Yang Lanlan, Gao Mingyu, Wang Chenning, Feng Dongjie, LV Xinrui. Research on facial expression recognition based on data enhancement [J]. Computer products and circulation, 2020 (11): 128-129

[14] Mehdi Mirza, Simon Osindero. Conditional generative adversarial nets. [DB/OL]. (2014-11-06). CoRR. abs/1411.1784

[15] Lu Chuanwei, sun Qun, Zhao Yunpeng, sun Shijie, Ma Jingzhen, Cheng mianmianmian, Li Yuanfu. A path extraction method based on conditional generative countermeasure network [J / OL]. Journal of Wuhan University (Information Science Edition): 1-10 [2020-11-10] https://doi.org/10.13203/j.whugis20190159.

Imperative to Build Network Security System and Speed Up the Future Network Construction

Wu Aiqun[1,2,3]

[1] Vice Director of Shanghai Committee Economy of CPPCC

[2] President of Shanghai Aerospace Information Technology Research Institute

[3] Vice President of Urban Risk Management Research Institute of Tongji University

Gao Zhangxing

ISO/IEC Future Network Standardization Expert, Science and Technology Consultant of Nanjing Future Science Technology City

***Abstract*—Network security is a comprehensive discipline involving computer science, network technology, communication technology, cryptography, information security technology, applied mathematics, number theory, information theory and other disciplines. Network security is to protect the hardware, software and data in the network system from accidental or malicious damage, change, and disclosure. The system runs continuously and reliably and normally, and network services are not interrupted. Although network security in the future market prospect is very good, but the Internet structural defects are known, because in the past nearly 20 years of rapid development of Internet technology, the pioneers of the IT industry seems to be technology focuses on the network flexibility and ignore the security of the network, so, in the site after the loss of a large number of network attacks, network security technology becomes emergent vacuum in the information technology, for the maintenance of computer network system security maintenance, inspection and repair network vulnerabilities, virus protection, organized by the national high strength against network has serious threat to all countries in the world, In such an environment, the core technology, key infrastructure is the only rely on independent innovation, create a new network system, with the new architecture, a new design, new technology, new resources, new standard and new application to open up a new network space, the building has an independent sovereign and independent control system of network security, it is imperative to accelerate the future network development.**

Keywords-New Architecture; New IP; New Technology; New Resources; New Standard; New Application

I. INTRODUCTION

The idea of "reorganizing the architecture of the web" is not a new idea. It has been around for 15 years. Since 2007, the international standardization organization ISO/IEC has been carrying out the Future Network standardization project of the new architecture Network system, and has set 2020 as the phased target for commercial use. This paper based on the research and development experience of ISO/IEC future network international standard, it shows that network system innovation is the core benefit of the development of China's information and communication technology, and it is imperative!

DOI: 10.21307/ijanmc-2020-040

II. THE STRUCTURAL FLAWS OF THE INTERNET ARE WELL KNOWN AROUND THE WORLD

China is no exception to the threat posed by state-level organized and high-intensity cyber confrontation to all countries in the world. In order to cope with state-level cyber confrontation, it is impossible not to change the situation that core technologies and key cyber infrastructure are in the hands of others. Core technologies cannot be bought and critical infrastructure cannot be sought. The only possibility is to rely on independent innovation, opening up a new network system, with the new architecture, a new design, new technology, new resources, new standard and new application space to open up a new network, the network construction of new frontier has sovereign and independent control of network security system, set up national network defense system, and for the enterprise and society to build a is not subject to sanctions and cyber threat to the survival and development space of peace. This is how countries and nations survive in the age of cyber warfare.

Under the situation that the Internet monopolizes the global information technology facilities, the proposal of the "new architecture network system" is bound to encounter opposition and obstruction from the Internet vested interests. The statement on Huawei NewIP by IETF, an American corporate standards agency, is a reflection of this. As the ancient saying goes, "each man is his own boss", which is a natural stance for the IETF. However, everything must have a reason. You can't object for the sake of objecting, you have to have a valid reason. Judging by the IETF's claims, the argument is flimsy.

The reason why Huawei proposed the "New IP" proposal emphasizes the structural defects of the Internet, and takes the 128-bit fixed-length address of the "next generation Internet" as an example to illustrate that in many application scenarios, shorter address length is needed, and the structure of the Internet does not meet the development needs of the society in the future.

It has long been universally acknowledged that the Internet is structurally flawed. Even many documents in the United States government say a lot about it. For example, at the beginning of the Internet design, it did not anticipate the tremendous changes and security threats brought by the development of science and technology decades later. It did not embed security into the architecture design, and many network security problems were caused by the structural defects of the Internet. If you were to enumerate the structural flaws of the Internet, the list could go on and on.

Taking 《Future Network Architecture and Its Security》 research report written by Chinese academy of sciences in 2014 as an example, makes a comprehensive analysis of the defects of Internet architecture from the perspective of security, and summarizes dozens of security architecture design requirements and solutions. Taking the ISO/IEC international standard draft of 《Future Network Security Architecture》 written by Chinese experts as an example, there are 100 technical indicators to be realized in the future network architecture design alone. These indicators correspond to the structural defects of the Internet one by one. If you include structural defects in other technical areas such as naming, addressing, routing, infrastructure, economics, topology, management, and so on, there are hundreds of structural problems that need to be addressed.

However, the study of the "new architecture of the network" has long since passed the stage of Internet defect studies and feasibility studies. So there is no need to devote too much energy to responding to the IETF's objections.

III. THE FUTURE OF THE INTERNET LEADS THE WORLD

In the past two decades, there have been two ideological trends in the development of the network technology system. One is the conservative approach stressed by Internet vested interests such as ISOC, ICANN, IETF and IANA that "the structural integrity of the Internet can only be maintained through gradual improvement". The biggest problem with this route is its inability to address structural flaws. Patch the method of "overlapping", security holes emerge in endlessly.

Another trend of thought has emerged since the beginning of this century, advocating a new approach, using the "empty cup design" method, a new blueprint for network architecture on a piece of paper, through a new architecture design to fundamentally solve the security problem. From China's ministry of information industry to establish a decimal network standard working group (2001), to the national science foundation GENI - FIND plan (2005-2006), to the future network international version of the ISO/IEC standardization project (2007), the ITU -t 13 working group (2008-2009), the future of the network, to the eu's "brad manifesto" (2008), the President of the United States national security telecommunications commission proposed "shot in the network security project" (2018), the Brics calls chairman Xi Jinping, To speed up the construction of "the Brics future network institute" (2019), and then to Huawei, China mail tunnels institute, China Unicom and China mobile "New IP" proposal, that a series of facts show that over the past two decades, the idea of network architecture reconstruction in China, the us, European and international standards organization has been a research hotspot and frontier technology in the field, has become an irresistible trend of the world.

In this world trend, the Internet standardization community will no longer hold a significant position. For Future Network (the Future Network, FN) as an example, the project is xi President called "the most authoritative international standardization organization" ISO/IEC organizations set up since 2007, currently has more than a dozen published planning technical report (ISO/IEC TR 29181 series), and are working on the Future Network architecture and protocol standard system (ISO/IEC 21558 and 21559 series standard).

As early as 10 years ago (2010), a member of a national body had written to ISO/IEC, claiming that the Internet standards were maintained by the IETF, and that the future network project violated the rights of the IETF, demanding that the project be stopped and withdrawn. On the basis of the position paper submitted by Chinese experts to ISO/IEC, ISO/IEC adopted the resolution that the future network is a completely new network system, which is not related to the Internet and does not fall within the scope of the authority of IETF, so there is no reason to stop or withdraw it. Subsequently, the TR 29181 series of technical reports, led by Chinese experts, received unanimous approval in a vote.

IV. THE NEW CYBER ARCHITECTURE WILL NOT HINDER GLOBAL CONNECTIVITY

Although the trend of reconstructing network architecture is irresistible, it still encounters great resistance and interference in the development process in the past. Whether it is international or domestic, Internet interest monopoly groups spread some wrong views through the media, misleading the decision-making and the public. If analyzed carefully, these views are all prejudices and fallacies, which simply cannot hold water.

For example, there is the "fragmentation" view that the new architecture of the network will lead to the fragmentation of the Internet and the "balkanization" of the Internet. But this view is groundless. Take the future network of ISO/IEC as an example. This is a brand new network system. It only considers the construction of its own system, but it will not touch the basic architecture and facilities of the Internet. The relationship between the future network and the Internet is like that between the new highway system and the old provincial and national highway system.

The construction of new highways will not hinder the survival and passage of old roads.

There is also a view that the new network will hinder globalisation and that countries will not be able to connect with each other. This, too, is a fallacy. Take the ISO/IEC future network as an example. It is a project organized by an international standardization body and actively participated by many countries in the world. It fully conforms to WTO norms and is recognized by the world. Not only developing countries will support it, but many developed countries are also optimistic about the project. In 2010, for example, the UK national committee submitted comments urging Chinese experts to submit technical proposals for future web naming and addressing as soon as possible. A telecom expert from the national association of France led the proposal to help China promote the new future network technology program to African countries. So, as long as the new network has clever design and application space, other countries will not be able to use it. It is an unreasonable assumption without any basis that other countries will not use it. There is ample evidence of this both in the historical literature of the international standards of the future web and in the previous summit declarations of Brics leaders.

There is a claim that, a new architecture of the network system will make the network investment benefit of telecom enterprises in the past suffered this view also belong to prejudice again in the future network, for example, France telecom has an expert in ISO/IEC has repeatedly stressed that the future network to consider good protect telecommunications enterprise's investment in 2009, Chinese experts to the ISO/IEC submitted a technical literature, in China a decimal network technology solutions, for example, suggests that the future network to ensure that the new network architecture independent complete and advanced nature, can also with the existing network connectivity, can protect the existing network investment Now that China Unicom and China mobile have joined Huawei's proposal, investment protection is no longer a concern.

Some people accuse China's independent innovation in the Internet system of "shutting the door on the outside world" or "narrow gauge train". This is typical idle talk and scaremongering. Take the future of the Internet as an example. It is an international standard. How can it become a "closed door"? In the future, the network international standard will have guidance and priority for adoption all over the world. Why is it a narrow-gauge train? The whole world has agreed on the future network planning scheme, almost all countries have the need for a new network architecture, how can it be impossible to gradually deploy and apply it globally? Moreover, the future network has already been designed to be compatible with the existing network and can be quickly deployed. Coupled with the advanced technology and full consideration of the application prospect, the future network has unlimited development potential.

V. NEW ARCHITECTURE FUTURE NETWORK RESEARCH AND DEVELOPMENT IN LINE WITH NATIONAL POLICY GUIDELINES

In Huawei's "New IP" proposal, it is clearly stated that this proposal belongs to the "future network" category. This enables Huawei's proposal to be strongly supported by future domestic network technology accumulation and national policies.

From the perspective of technology accumulation, China is the first country in the world to carry out the research on the new architecture network system. As early as the late 1990s, China has started the pre-research and tackling of the new network system, and has made technological breakthroughs and obtained patent and copyright protection. In 2001, the ministry of information industry of China set up the working group on decimal network standard, and soon promulgated the industry standard of "digital domain name specification" (2002). In 2004, when the

ISO/IEC JTC 1 / SC 6 Xi'an plenary session considered a future network standardization project with an entirely new architecture, China voted in favor. In the following ten years, China's national members have contributed a lot of technical documents to the future international network standards. The international standards committee, the ministry of industry and information technology and the China institute of electronic standardization have held several meetings to promote the future network standardization, and the central leadership has issued important instructions on many occasions. China is a major contributor to the naming and addressing and security solutions in the future network core technology areas. China voted in favor of the future of Internet international standards. Therefore, it is the national position of the People's Republic of China to establish a new architecture network system based on international standards. This position admits of no challenge.

In terms of domestic policy, the Chinese government has always attached great importance to the future research and development of network technology. As early as 2013, the state council announced in document no. 8 that the gradual improvement of the Internet based on TCP/IP could no longer meet the needs, and that it was necessary to break through the basic theory of the future network, build future network experimental facilities, and incorporate the future network into the national medium - and long-term science and technology planning. In 2015, after a year-long investigation, the Chinese academy of sciences submitted a report to the state council, recommending the establishment of major national projects to promote future network research and development with the will of the state. In 2017, the general office of the CPC central committee, the general office of the state council, the cyberspace administration of the CPC central committee, the state standards commission, the ministry of science and technology, and the commission of science and technology of the central military commission all released documents, listing the future network as one of the key technology areas that are "forward-looking, subversive and killer" during the 13th five-year plan period. In 2019, at the brics informal summit in Osaka, Japan, President Xi Jinping proposed to speed up practical projects such as the brics future network institute.

In such a situation, our country's scientific research institution should keep highly consistent with the party central committee and the State Council, the propaganda department and the mainstream media should also be clear to defend persevere in the position of network sovereignty, avoid becoming the Internet interests abroad monopoly the mouthpiece of the group to maintain its backward system, not for our country in the future, a new architecture of sovereignty for the construction of network system.

VI. IT IS URGENT TO REBUILD THE NEW SYSTEM AND ACCELERATE THE SOCIAL BUSINESS APPLICATION

President Xi Jinping has always stressed that core technologies cannot be acquired. We need to adhere to independent innovation to change the situation that core technologies in the field of information are controlled by others. In his remarks during the 2019 cyber security publicity week, Xi Jinping called for equal emphasis on governance and innovation in cyber security. In his speech at Davos 2019, vice chairman Wang mentioned innovation seven times. "We can only find ways to better slice the cake as we make it bigger," and said. "we can't stop and argue endlessly about how to cut the cake. Shifting the blame to others will not solve the problem.

We now can completely don't have to dwell on the right and wrong of the Internet, and should make full use of the new infrastructure network system reconstruction strategic opportunities brought by the international trends, build a new architecture of the network system of big cake, establish complete sovereignty of network new space for the future in our country, and then lift force of the country, the construction of new cyberspace into an exempt from cyber threats, with endless resources, people's safety work and future generations of survival in the new world, new frontiers and new paradise. This will be a great cause in the present and future. The use of innovation to develop the network technology system is a sovereign state's right to survival and development, no country has the right to interfere.

The urgent task of advancing this cause is to accelerate the development of international standards for the future network. Huawei's dispute with the IETF over "New IP" is further proof of the importance of international standards. We are developing a new architecture not just to protect ourselves, but to address the urgent need of people around the world for equal sovereignty and a secure network. International standards are not only a platform for technical exchanges among countries on the new network technology system, but also a bridge for China's future network system plan leading the research and development to the world. Although China has made proud achievements in this field, there is still much work to be done to form a complete system of future network technology standards. In particular, the future network security architecture based on the new design of the international standard scheme is the key factor of the future network success or failure is the future network crown jewel. In this field, China has achieved world-leading results, and more national resources are needed to integrate it into the future network international standard system. In the environment of increasingly fierce competition for international standards in network system, enterprises will lose precious opportunities if they are allowed to face the competition of "full government power" from other countries.

At the same time, China should urgently promote the practical deployment of the new architecture of the future network technology system. The expected commercial time of future Internet international standard is 2020. Because our country starts in this domain early, already had the ability that invests commercialize now. This preparation includes not only standards and equipment, but also application scenario design. The Internet of things is the biggest application scenario of the future network. The Internet of things based on the new future network architecture has been issued by the ministry of commerce and the ministry of industry and information technology in 2002, 2010 and 2016 respectively. Due to the advantage of being a late starter, China's future Internet of things technology is more in line with the application of the Internet of things. From the perspective of social research, there is a very urgent and widespread need.

In the increasingly severe international situation and the increasingly imminent threat of cyber warfare, accelerating the deployment of China's autonomous and controllable future network is no longer an option, but a necessary option. The spread of covid-19 in the us and Europe has led to growing calls from western anti-china forces for China and the us to decoupled, a trend that is bound to spread to the cyber sector. We need to consider the question: what if the network decouples? Are we ready? This problem is not only Huawei will face, but also the difficult problem that people all over the country can not avoid.

VII. CONCLUSION

The "structural flaws in the Internet" proposed by Huawei in its "New IP" international standard proposal is a universally recognized fact that cannot be changed even if the IETF denies it. The proposal was submitted to the international standards organization. The IETF has comments that can be submitted to international institutions through its national membership. Huawei's idea and proposition of "rebuilding the network architecture" fully conform to the position of ISO/IEC international network standards in the future, which has been repeatedly demonstrated for 13 years, and its rationality and feasibility have been unanimously approved by the international community. The IETF's position reflects its desire to protect the interests of the Internet, but whether it meets the needs of people around the world needs to be evaluated and considered in international standards bodies. The ISO/IEC one country, one vote decision-making mechanism is the best mechanism to ensure the sovereign equality of all countries and the fairness of the world.

The route of "rebuilding network security architecture" represented by ISO/IEC future network international standard has been the trend of the world, representing the most important field and direction of the future development of information and communication technology in the world, and has reached the stage of commercial construction. The future network is a frontier technology field which the Chinese government attaches great importance to. Its technical scheme has been recognized by the world and has a wide application prospect in the world. The future network will sui generis, don't have to rely on existing network infrastructure support, but for the rapid deployment and application, the design with the existing network compatibility mechanism, will not affect the structure stability of the Internet, not push the fragmentation of the Internet, will not endanger the telecom enterprises existing investments, will not lead to a "closed" or "narrow gauge train" phenomenon. In the future, the network also has new core resources with great industrial value, which can drive huge social and economic value and industrial development space.

"Ten years to sharpen a sword". The future of our country network is twenty years, in the theoretical study, the top design, the international standard, the system scheme, the security architecture, autonomous control, the core equipment, sovereign legislation, offensive and defensive drills, and strategic planning, etc, have mature solution, prepared and available equipment system, can be said to be the everything is ok, only owe the east wind. The investment of Huawei and other organizations indicates that the future network is on the fast track of development.

At present, the international situation is very serious, and the threat of national-level cyber confrontation and cyber warfare is increasingly imminent. It is an urgent task to strengthen the construction of the national network security defense system with the new architecture of the future network technology system. In the past, there have long been differences in the development path of the network system, but there has been a growing consensus to rely on the new architecture of the network to strengthen national cyber security. We should seize the opportunity, give priority to national and national interests, discard past grievances, unite all forces that can be united, form the broadest possible united front, and accelerate the construction of a new architecture for future network standardization and social and practical deployment.

ABOUT THE AUTHOR

Wu Aiqun: Deputy director of the committee of economy of the CPPCC committee of Shanghai, President of Shanghai academy of aerospace information technology, vice President, professor, doctoral supervisor of urban risk management institute of tongji university, etc. Telephone number: 13801987952.

Gao Zhangxing: ISO/IEC future network standardization expert, science and technology consultant of nanjing future science and technology city.

Editor in charge: Yang Yining

REFERENCES

[1] Jung H., & Koh S.J. Mobile Optimized Future Internet", http://protocol.knu.ac.kr/tech/CPL-TR- 10-01-MOFI-12.pdf, Jun. 2010Electronic Industry Standard of the People's Republic of China [M]. 2002.

[2] Xie Jianping, HUANG Changfu. Current Situation and Development of Decimal Network [J]. Information Technology and Standardization,2004(04):5-9.

[3] Zhang Yunghui, Jiang Xinhua, Lin Zhangxi. Comparison between IPv6 and IPv9 [J]. Computer Engineering,2006(04):116-118.

[4] Soliman H., Castelluccia C., Malki K. E., Bellier L. Hierarchical Mobile IPv6 (HMIPv6) Mobility Management," IETF RFC 5380, Oct. 2008