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

Cooperate:

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

Approval:

Library of Congress of the United States

Shaanxi provincial Bureau of press, Publication, Radio and Television

Address:

4525 Downs Drive, St. Joseph, MO64507, USA

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

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

Website: www.ijanmc.org

E-mail: ijanmc@ijanmc.org

xxwlcn@163.com

ISSN: 2470-8038

Print No. (China): 61-94101

Publication Date: December 28, 2022

Associate Editors

Table of Contents

Remote Sensing Image Object Detection Method Based On Improved YOLOv3

Zhiyuan Lu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: luzhiyuan@st.xatu.edu.cn

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

***Abstract*—In order to solve the problem that irregular targets and dense targets are difficult to be detected in optical remote sensing images, this paper improved the YOLOV3 model Firstly, in order to further combine the feature information of different scales, the PaNet structure is introduced into the FPN part of the original YOLOv3, and the obtained effective feature layer is continued to be extracted for a round of feature. The feature is not only up-sampled to achieve feature fusion, but also down-sampled again to achieve enhanced feature fusion SimOTA method is introduced to dynamically match positive samples and set different positive sample numbers for different targets, which not only improves the speed of the algorithm, but also reduces the extra hyperparameters Experimental verification using richer DOIR data sets shows that the detection ability of the improved algorithm is significantly improved. Compared with the original YOLOv3, its mAP improves by 15.1 points, among which the detection accuracy of dense small targets is improved the most.**

Keywords-Remote Sensing Image; Object Detection; Yolov3; DOIR Dataset

I. INTRODUCTION

Object detection is an important computer vision task that deals with detecting instances of visual objects of a certain class (such as humans, animals, or cars) in digital images. The objective of object detection is to develop computational models and techniques that provide one of the most basic pieces of information needed by computer vision applications: What objects are where? [1]. Up to now, many excellent object detection algorithms have been proposed to classify them according to time. Object detection methods can be divided into traditional object detection method and object detection method based on deep learning. The traditional object detection methods include HOG [2] detector and DPM [3] detector Object detection methods based on deep learning can be divided into two-stage detection methods and single-stage detection methods. Typical representatives of two-stage object detection methods are RCNN [4], Fast RCNN [5] and Faster RCNN [6] and typical representatives of single-stage target detection methods are YOLO [7][8],[9]. SSD [10] and RetinaNet [11] etc.

In the early days when remote sensing technology was born, remote sensing image classification and recognition adopted manual visual and manual marking methods [12]. With the development of deep learning technology, more and more excellent object detection algorithms are applied to remote sensing image object detection. But due to the optical remote sensing image of object and the natural scene in the image difference is very big, mainly embodied in the following aspects: the target scale diversification, different categories of object scale is large, the same category object scales under different shoot height difference is bigger also Image of the upper vertical view shows only the object. Small object with irregular shape object accounts for more than, in the background of more complex scenarios, target and background is easy to confuse Target due to the cause of the shooting Angle in the direction of the arbitrary to these problems such as optical remote sensing image object detection and recognition of enormous challenge [13]. Therefore,

DOI: 10.2478/ijanmc-2022-0031

the problem of remote sensing image object detection is a very challenging topic.

In view of the above difficulties in remote sensing image target detection, an improved algorithm based on YOLOv3 is proposed for remote sensing image object detection. By introducing PaNet [14] method to enhance feature fusion, features of different scales are further fused to deal with the problem of target scale diversity and large number of small objects. SimOTA [15] method was introduced to carry out more precise screening of positive samples with less computation and parameter number. In order to improve the detection accuracy of the original YOLOv3 algorithm detection speed.

II. INTRODUCTION OF YOLOV3 ALGORITHM

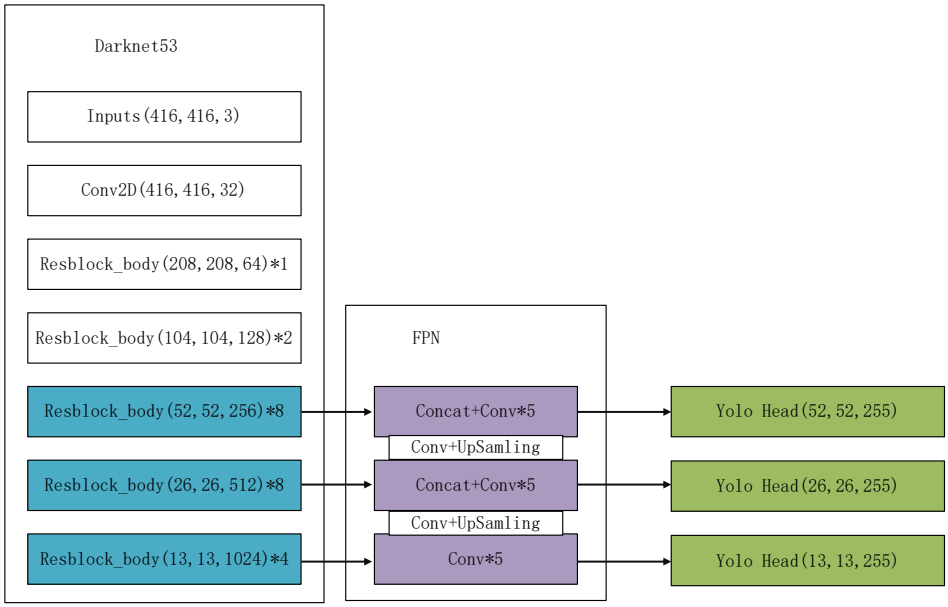


Figure1. YOLOv3 overall structure

YOLOv3, the third version of YOLO algorithm, was first proposed by Joseph Redmon, who creatively treated the object detection problem as a regression problem. Yolov3 network structure as shown in Fig. 1 It first adjust the image size is 416×416, the input to the network, after a series of convolution of image feature extraction, and then through a pyramid fusion of different scale layer characteristics, characteristics of results can be divided into three categories, respectively used to predict the small size targets The goal of medium size and the goal of large size.

A. Backbone Network of YOLOv3

The backbone network of YOLOv3 adopts Darknet-53 network, in which the average pooling layer, full connection layer and Softmax are removed, and is mainly composed of convolution and residual modules. An important feature of Darknet-53 is that Residual network is used. Residual convolution in Darknet-53 is carried out first with a convolution kernel size of 3×3, the convolution with step size 2 will compress the width and height of the input feature layer. At this time, a feature layer can be obtained. After the feature layer is named layer, the 1×1 convolution and 3×3 convolution are carried out for the feature layer, and the result is added to the layer, and then the residual structure is formed Through constant 1×1 convolution and 3×3 convolution and superposition of residual edge, can greatly deepened the characteristics of the residual network is easy to optimize, and to improve accuracy to by adding considerable depth. Its internal residual block uses the jump connection, easing in the depth of neural network to increase the depth of gradient disappeared.

Each convolution part of Darknet-53 uses the unique DarknetConv2D structure. L2 regularization is performed during each

convolution, and Batch Normalization is performed after completion of convolution with Leaky ReLU. The normal ReLU is to set all negative values to zero, Leaky ReLU gives all negative values a non-zero slope which can be mathematically expressed as:

$$y_i = \begin{cases} x_i & x_i \geq 0 \\ \frac{x_i}{a_i} & x_i < 0 \end{cases} \quad (1)$$

B. The features are constructed as predicted results

The process of obtaining prediction results from features can be divided into two parts: building FPN [16] feature pyramid to enhance feature extraction; Three effective feature layers were predicted by YOLO Head.

- FPN feature pyramid was constructed to enhance feature extraction.

In the feature utilization part, YOLOv3 extracts multiple feature layers for target detection. There are altogether three feature layers extracted. Three feature layers are located in different positions of main Darknet-53, namely middle layer, middle and lower layer, and bottom layer. The shapes are (52,52,256), (26,26,512), (13,13,1024). After obtaining three effective feature layers, the three effective feature layers can be used to construct the FPN layer. The construction method is as follows: the feature layer of 13×13×1024 is convolution processed for 5 times. After processing, YOLO Head is used to obtain the prediction results, and part of it is used for up sampling UpSampling2d and 26×26×512 features. The shape of the combined characteristic layer is (26,26,768). The convolution processing is performed again for 5 times in combination with the feature layer. After the processing, YOLO Head is used to obtain the prediction results, and part of up sampling UpSampling2d is used to combine with 52×52×256 feature layer. The shape of the combined feature layer is (52,52,384). Five convolution processes were carried out in combination with the feature layers. After the processing, the prediction result of feature pyramid was obtained by YOLO Head. Feature fusion of feature layers of different shapes was beneficial to extract better features.

- The prediction results were obtained by YOLO Head.

Three enhanced features can be obtained by using FPN feature pyramid, whose shapes are (13,13,512), (26,26,256), (52,52,128), and then the feature layers of these three shapes are passed into YOLO Head to obtain the prediction result YOLO Head is essentially a 3×3 convolution plus a 1x1 convolution, 3×3 convolution is for feature integration, 1×1 convolution is for adjusting the number of channels.

The shape of the output layer is (13,13,75), (26,26,75), and (52,52,75). The last dimension is 75 because the graph is based on VOC dataset. It has 20 classes, and YOLOv3 is for each feature layer There are 3 prior boxes for each feature point, so the number of channels for the prediction result is 3×25. If coco training set is used, there are 80 kinds of classes, and the last dimension should be 255=3×85. Shape of the three feature layers is (13,13,255), (26,26,255), (52,52,255). The actual situation is that N×416×416 images are input, and three shapes of (N,13,13,255), (N,26,26,255) and (N,52,52,255) will be output after multi-layer operation, corresponding to each figure is divided into 13×13, 26×26 and 52×52 grid Position of 3 prior boxes on.

C. Decoding of prediction results

According to the second step, the prediction results of three feature layers can be obtained, and shapes are: (N,13,13,255), (N,26,26,255), (N,52,52,255), Each effective feature layer divides the whole image into grids corresponding to its length and width. For example, the feature layer (N,13,13,255) divides the whole image into 13×13 grids. Then several prior boxes are built from each grid center, these boxes are pre-configured boxes for the network, the network's predictions determine whether these boxes contain objects and what kind of object it is. Since each grid point has three prior boxes, the prediction result above can be reshape into: (N,13,13,85), (N,26,26,3,85), (N,52,52,3,85). Where 85 can be split into 4+1+80,

where 4 represents x_offset, y_offset, h and w the four, the parameters of the prior box 1 represents a priori whether the box contains objects, 80 represent the types of the a priori box, because coco got the 80 class, so there is 80 but this result does not correspond to the final forecasting box in the picture, you also need to decode can complete YOLOv3 decoding process is divided into two steps:

- Add x_offset and y_offset to each grid point, and the result is the center of the prediction box.
- Then, the width and height of the prediction frame can be calculated by combining the prior frame and h and w, so that the position of the whole prediction frame can be obtained. After obtaining the final prediction results, score sorting and non - maximum suppression screening should be carried out.

III. IMPROVED YOLOV3

A. Strengthening feature fusion

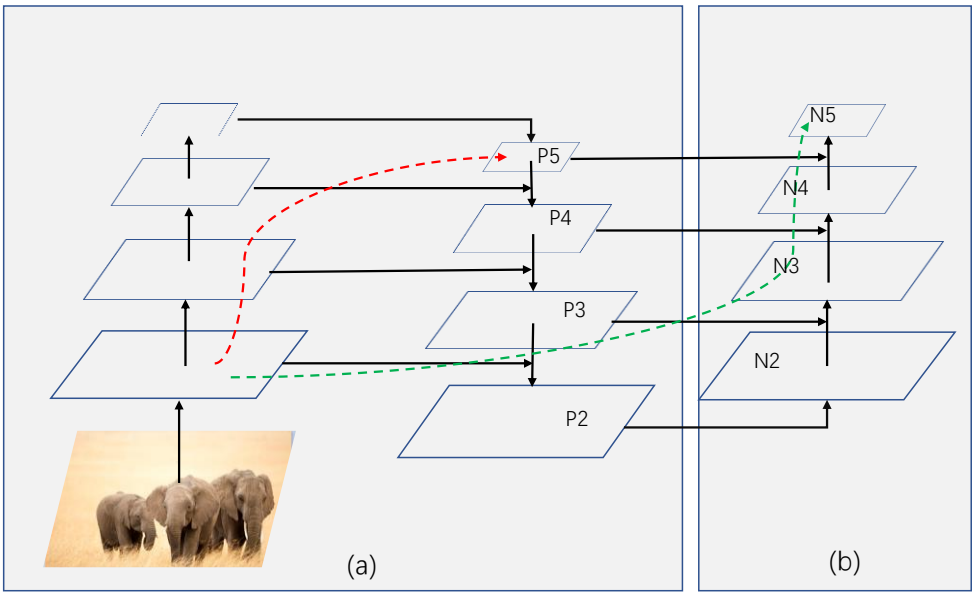


Figure2. PaNet overall structure

PaNet(Path Aggregation Network) is the addition of a bottom-up Path Augmentation to FPN. You can see in Figure 2 (a) the FPN is top-down route, through the lateral connection, transfer the strong semantic characteristics of the top down, only to enhance the characteristics of the pyramid semantic information For example, when the underlying characteristics to the P5 (red line), after very multilayer networks among them, at this point at the bottom of the target information is very fuzzy, so the FPN extensions, added a bottom-up route (green route, bottom ->P2-> N2~N5, where the path passes through less than 10 layers), as shown in Fig. 2 (b), thus compensating and reinforcing the positioning information.

By applying PaNet to YOLOv3, another round of down sampling is conducted on the basis of its FPN structure to Further strengthen feature fusion, as shown in Fig. 3, thus improving the accuracy of YOLOV3 for detecting dense small targets and multi-scale targets in remote sensing images.

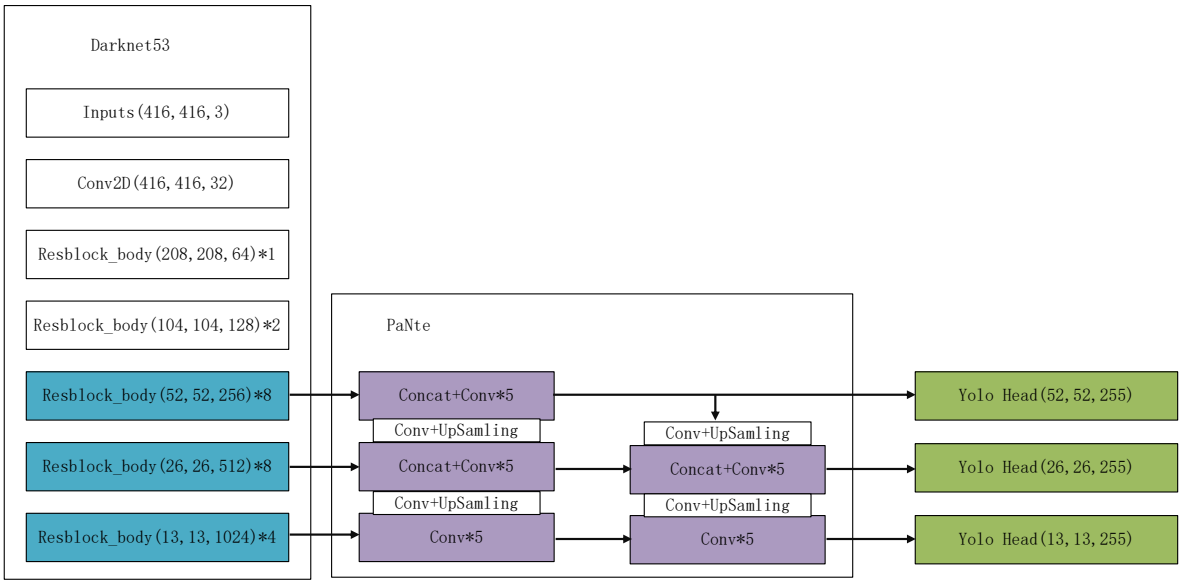


Figure3. YOLOv3 structure after introduction of PaNet

B. SimOTA method

SimOTA's function is to set different positive sample numbers for different targets, such as ants and watermelon. Traditional positive sample allocation schemes usually assign the same positive sample numbers to watermelon and ants in the same scene, so either ants have many low-quality positive samples, or watermelon only has one or two positive samples. It's not appropriate for either way of distribution. Therefore, SimOTA firstly calculates a cost matrix, which represents the cost relationship between each real box and each feature point. The cost matrix consists of three parts:

- The higher the degree of overlap between each real box and the prediction box of the current feature point, it means that this feature point has tried to fit the real box, so its Cost will be smaller.
- The higher the type prediction accuracy of each real box and the current feature point prediction box is, it also means that this feature point has tried to fit the real box, so its Cost will be smaller.
- Does the center of each real frame fall within a certain radius of the feature point? If the center of each real frame falls within a certain radius of the feature point, it means that the feature point should fit the real frame, so its Cost will be smaller.

The process application of SimOTA is as follows:

- Calculate the coincidence degree between each real box and the current feature point prediction box.
- Calculate the IOU of the ten prediction frames with the highest coincidence degree and the real frame to get the K of each real frame, which means that each real frame has K feature points corresponding to it.
- Calculate the type prediction accuracy of each real box and the current feature point prediction box.
- Judge whether the center of the real box falls within a certain radius of the feature point.
- Compute the cost matrix.
- The k points with the lowest cost are taken as positive samples of the real box.

IV. EXPERIMENTAL RESULTS ANS ANALYSIS

A. Experiment preparation

All experiments were carried out on a Linux operating system computer equipped with AMD

R9-5900HX CPU, Nvidia 2080Ti GPU, 16G video memory and 16G memory. The remote sensing image dataset used in the experiment is DOIR dataset, which contains 23463 remote sensing images and 192472 object instances. These instances are manually marked as boundary boxes of axial pairs, covering 20 common object categories. The size of data set images is 800×800 pixels, and the spatial resolution is 0.5m~30m Among them, the training set accounted for 1/2, the verification set accounted for 1/6, and the test set accounted for 1/3.

B. Evaluation indicator

The evaluation indicators of the experiment include precision(P), recall (R), average precision (AP), and mean Average Precision of all categories(mAP). The calculation formula of each index is as follows:

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$AP = \int_0^1 P_{smooth}(r) d_r \tag{4}$$

$$mAP = \frac{\sum_{j=1}^{K} AP_i}{K} \tag{5}$$

Where TP, FP, TN and FN represent positive samples with correct predictions and positive samples but wrong predictions and negative samples with correct predictions and negative samples with wrong predictions. $P_{smooth}(r)$ is a smooth P-R curve.

C. Comparison of experiment results

The comparison results of YOLOv3 and the improved YOLOv3 in this paper are shown in TABLEI. It can be seen that the detection accuracy of the improved YOLOv3 for small objects has been significantly improved, and the mAP has increased by 15.1 points. In terms of single category, the improved method is applied to the Wind mill, Airplane, Tennis Court, Expressway service area, Baseball fields and other types of small object detection performed better. The best-performing ballpark had an accuracy of 98.2 percent.

TABLE I. DETECTION ACCURACY OF EACH CATEGORY ON DOIR DATASET

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
Airplane	Airport	Baseball field	Basketball court	Bridge	Chimney	Dam	Expressway service area	Expressway toll station	Golf course
c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
Ground track field	Harbor	Overpass	Ship	Stadium	Storage tank	Tennis court	Train station	Vechile	Wind mill

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c16	c17	c18	c19	c20	mAP
YOLOv3	90.9	69.5	81.7	78.6	61.2	69.7	66.9	88.6	74.4	61.1	89.1	44.9	49.7	90.4	70.6	68.7	87.3	59.4	68.3	78.7	72.5
改进 YOLOv3	96.7	95.1	98.2	85.6	71.7	95.4	79	96.1	91.3	90.1	92.4	68.2	74.2	92.5	90.8	80.5	96.7	77	84.9	96	87.6

P-R curves of each category are shown in Fig. 4, and the detection effect of the algorithm in this paper on small targets in remote sensing images is shown in Fig. 5, from which it can be seen that the improved YOLOv3 has a better detection effect on small targets.

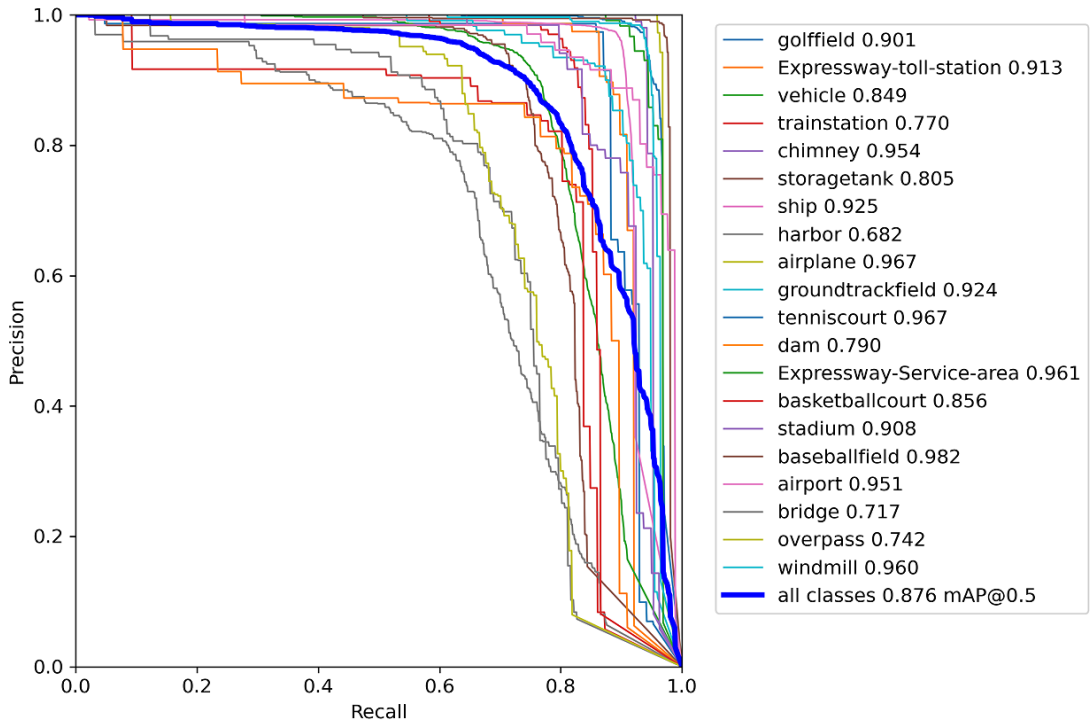


Figure4. P_R curve

(a) (b)

Figure5. The detection effect of improved YOLOv3

V. CONCLUSION

Based on remote sensing image inherent in the small object proportion is high, the object scale inconsistent problems, puts forward the improved YOLOv3 algorithm used for target detection in remote sensing image, the original structure of the characteristics of the pyramid module on the basis of round of sampling was conducted again, to strengthen the feature extraction and improved its ability to cope with multi-scale object In general, the improved YOLOv3 algorithm solves the problem of low detection accuracy of partially

dense small targets in previous algorithms, and improves mAP by 15.1 points.

REFERENCES

[1] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[J]. arXiv preprint arXiv:1905.05055, 2019.

[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[3] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE conference on computer vision and pattern recognition. Ieee, 2008: 1-8.

[4] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[5] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[6] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv:1804.02767, 2018.

[10] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[11] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[12] Tuia D, Persello C, Bruzzone L. Domain adaptation for the classification of remote sensing data: An overview of recent advances[J]. IEEE geoscience and remote sensing magazine, 2016, 4(2): 41-57.

[13] Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: A survey and a new benchmark[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307.

[14] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[15] Ge Z, Liu S, Li Z, et al. Ota: Optimal transport assignment for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 303-312.

[16] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

3D Face Feature Processing and Recognition Technology

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

***Abstract*—Nowadays, we frequently use a lot of biometric identification technology means, such as: iris recognition, voice recognition and face recognition, but compared with other human biometric features, face information is easier to collect and more easily accepted by the public. According to the survey, in interpersonal contact, most people get information based on the first impression, even if there are changes in external conditions such as age, expression, makeup, etc., but the brain can still recognize these information, this phenomenon shows that face does have more feature information. From the information of a person's facial features, we can effectively distinguish each person. Face recognition is a technology that analyzes the face data and then uses it as an authentication mechanism. Generally, cameras are used to track and collect human faces, model theory and then analyze the collected binary data stream to extract relevant feature data. The research of human face recognition technology began in the middle of 1960s. After 1980s, due to the development of computer technology and the improvement of optical image processing technology, it gradually matured. The real use stage of face recognition is in the late 1990s, with Germany, Japan and the United States as the main technologies.**

Keywords-Face Recognition; Model Theory; Image Processing

I. INTRODUCTION

The realization of face recognition technology has three steps, the first step, collect the image of the face as a sample, the larger scientific research institutions have their own face database, the second step, extract the need to process the key information of the face, the last step is to compare with their own face database. Face recognition technology has made important contributions in the fields of psychology, criminology, medicine and so on. China began to put into the research of this technology late, but still has made outstanding results. In the later stage, another face recognition processing technology was proposed by Blanz et al. This method simulates the 3D face image space, and then matches the 3D face model with the image space, so as to estimate the shape and texture information of the face. This algorithm has a high recognition capability, the recognition rate is up to 88 % [1]The research methods that need to be adopted for face information usually rely on these four specific operation steps to achieve: The first step is face localization and detection: For a given image information, first of all, whether it is belong to face images or contain face, if contain face, defining the face area in the image, then extracted face part, because there is much easier to change the face itself factors, and is strongly influenced by the external environment, so in order to realize the recognition of human faces and detect need quite mature technology, Otherwise, it is difficult to achieve efficient face recognition. The second step is facial feature extraction: usually after the

machine determines the image as a face, it will use some algorithm to extract the facial feature data, and then save it to the face database. The most important algorithms are surface and algebraic features. The third step is the real stage of recognition: will obtain the face image and the database exists in the template to match, by comparing the same and different places to get the correct recognition results. This stage is almost not affected by external factors, mainly to see whether the algorithm is appropriate, can effectively identify, and obtain the correct face feature information. The fourth step is external interference project analysis: summarize and analyze the external environmental impact. The algorithm is improved to overcome the influence of these objective factors [2].

II. RESEARCH METHODS OF FACE RECOGNITION

Template method is the most classic pattern recognition method to solve the problem of face recognition, this method repeatedly uses the characteristics of face texture and different gray values, the main principle of this method is to identify the face image and all the templates in the database to find the best matching effect of the face. Because this method has great requirements on the image template, the two templates need to have the same proportion, the same face orientation and the same lighting conditions, so the ratio normalization and gray value normalization of the two templates are needed in the preprocessing stage. The most easy to implement method is to face image approximation as oval, then, in recognition of the image to find elliptic, another method is a set of templates to represent the whole face template, this template including small template features of various organs of the human face, but this approach is more demanding, must use each feature of different edge profile, The traditional method is based on edge extraction, so it is difficult to obtain continuous edge data. Even if relatively continuous and reliable data are obtained, it is difficult to automatically extract the eigenvalues of different small templates that we need. Later, variable templates were used to extract the edge information of different templates in template group. The variable template consists of setting a set of variable characteristic parameters according to the existing knowledge, which depends on the special value of the energy function. Firstly, the prior knowledge of boundary value, peak value, trough value, trough information and image shape information is used to design the function, and then the parameters are modified in the direction of energy function reduction. At a minimum, the template shape corresponding to this set of parameter locks is most consistent with the feature shape. Although the recognition method based on feature setting has faster recognition speed and smaller memory consumption, the method based on elastic template is better than the traditional template method in recognition rate and other factors. Therefore, some people specially compare the recognition method based on facial geometric features and the recognition method based on elastic template. The result shows that elastic template method belongs to geometric feature method.

The essence of feature face method is realized by K-L transform, which is the best orthogonal transform method for image compression in image processing technology. Kirby and Sirovich successfully described the method of face image information through component analysis. They calculated the coordinate system with the best image compression rate for the original image information. Each coordinate is actually the eigenpoint coordinate system that they define. [4] They argue that theoretically, any face image template can be reconstructed from two sets of

images, one is the image of each face, and the other is the standard template image, also known as the eigenimage. The number of repetitions can be obtained by mapping the face image with the corresponding feature image by projection. Turk and Ponte Land argued that a great deal of facial information could be obtained if a set of feature images could be weighted and reconstructed. The effective recognition method is to identify the human image by analyzing the complex features of the image and comparing the feature weights needed to reconstruct the image. Thus, each face can be represented by a set of reconstructed weights. In short, vectors of higher dimensions are projected into the region of vectors of lower dimensions by matrices with specific characteristics. The eigenmatrix is characterized by a low-dimensional vector and retains the main information. That is, the corresponding high-dimensional vector can be completely reconstructed through the state eigenvector and eigenvector matrix. Compared with the image itself, this expression can completely reconstruct the corresponding high-dimensional vector. This expression is a highly compressed form relative to the image itself [5][6]. Artificial Neural Network (ANN) has a good ability to classify complex patterns, which leads to its widespread use in face recognition. There are many different models of neural network models, each of which describe and simulates different levels of biological nervous system from different perspectives. Representative neural network models include perceptron, multilayer mapping BP network, RBF network, Hopfield model and so on. At present, in the practical application of artificial neural network, most of the neural network models are BP network and its variants. It is the most important part of artificial neural network [7].

III. Classification of Algorithms

There are three methods to detect facial feature points: the traditional method based on active shape model (ASM) and active appearance model (AAM), the method based on cascade shape regression, and the method based on deep learning. When the method is classified by whether there are parameters, ASM, AAM and CLM are parametric methods, while cascade regression and deep learning are non-parametric methods. Methods based on parametric shape model can be divided into local method and global method according to different shape models. Non-parametric methods can be subdivided into sample-based methods, graph-based methods and cascaded regression methods. And a deep learning-based approach. See the literature for a more detailed division. Currently, the most widely used and accurate method is based on deep learning. Therefore, in this paper, we mainly study the application of deep learning in face important point detection. Then, referring to the comprehensive review of face detail detection, two main face detection methods are proposed: parametric and non-parametric. This classification method is more acceptable and can make people better understand the meaning of the existence of parameters.

Figure 1 shows the classification figure:

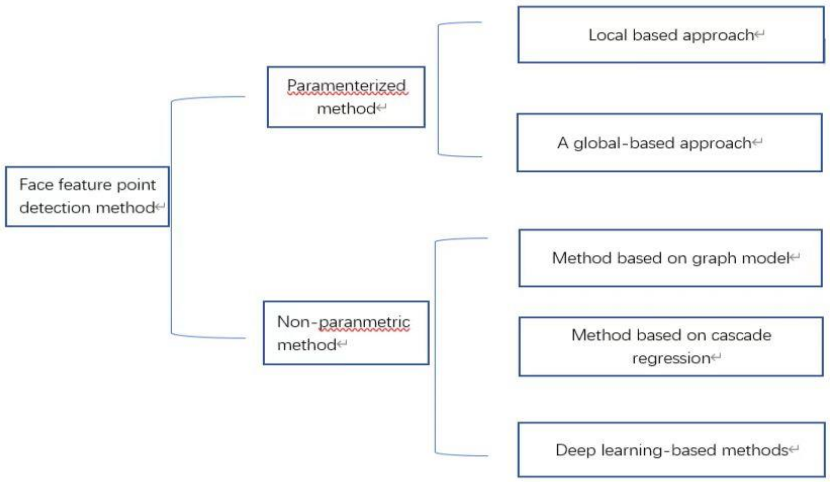


Figure 1. Classification diagram

A parametric model is a data distribution of interest to a particular target, such as a Gaussian model or Gaussian mixture Dell. Methods based on nonparametric models do not meet the requirements fundamentally, and their data should be specified to meet the requirements on the basis of distribution. Main models or non-parametric models (former) have fixed parameters, the latter numbers of parameters are Elites parameters Bakelites, Hikaru, have Hikaru electronic games with grades (Manufacture Kute STAB Silicone S Factory Yip electronic games, Hikaru of Hikaru Nad and Hikaru with Corning. Non-parametric models with video Hikaru are graph-based methods, cascaded regression based methods, deep learning and other methods [8].

The basic standard at this stage is to see the algorithm to obtain the face feature points and the actual face point between the deviations. In the quantitative evaluation of the deviation, because the face image is different, the actual size of the shooting Angle and distance will be different, in order to facilitate the comparison of the performance of the algorithm under the same conditions, the method of data normalization should be used to maintain the constant factors. At present, the main reference method is to rely on the method of judging the distance between the two eyes for face size standardization.

A. Common database

Active Database: CMU Multi-PIE Face Database, formed by collecting multiple face data at four conferences between October 2004 and March 2005, supports the development of face recognition algorithms in a variety of conditions, such as posture, lighting conditions and expression. The database contains 337 topics and over 750,000 images of 305GB of data. Six different expressions were recorded: expressionless, smiling, surprised, leering, disgusted and screaming. Objects were recorded in 15 views and 19 different lighting conditions. A subset of the database is labeled with 68 or 39 points. The XM2VTS database collects 2360 color images, sound files and 295 people's 3D facial models. The 2360 color images have 68 key points. The AR database contains more than 4,000 color images of the faces of 126 people (70 men and 56 women). With all the uncontrollable conditions in place, Ding and Martinez manually

labeled 130 key points for each face image. The IMM database contains 240 color images of 40 people (7 women and 33 men). Each image is marked with eyebrows, eyes, nose, mouth and chin, a total of 58 markers. The MUCT database contains 3755 images of 276 people, each with 76 keypoints. The faces in the database were photographed under different lights, different ages and different races.

The PUT database collected 10,091 high-resolution images (2048×1536) of 10,091 individuals with partially controlled lighting conditions and rotation along pitch and yaw angles. Each image is labeled with 30 key points.

B. Face++ version DCNN

This paper uses Face++ Megvii's face recognition API. The API algorithm is improved on the traditional DCNN model, and a coarse-grained to fine-grained human face key point detection method is proposed, which mainly solves the problem of high-precision positioning of 68 human face key points. The algorithm divides the key points of the face into internal key points and external contour key points. Inside the face is not the main point eyebrows, eyes, nose, mouth a total of 51 key points, contour key points contain 17 key points. For internal keypoints and external keypoints, the algorithm uses two cascaded CNNS for keypoint detection in parallel. The network structure is shown in Figure 2 below.

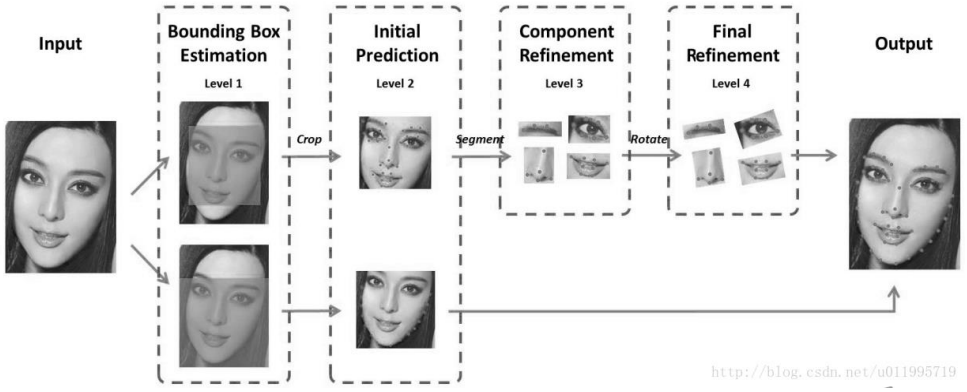


Figure 2. Network structure

For internal key points, a cascade network of four levels is adopted to continue the detection. Among them, the first layer is mainly used to locate the boundary of facial organs. The second layer is the position where the output predicts the 51 key points, which is a coarse-grained positioning for the purpose of initialization for the third layer. The third layer will be based on the features of different wonders of the face from coarse-grained to fine-grained positioning; The input of the fourth layer is to rotate the output of the third layer to a certain extent, and finally output the positions of 51 key points. For the 17 external key points, only a two-layer cascading network is used for detection. The first layer plays the same role as the internal key point detection, mainly to obtain the bounding box of the face contour; The second layer directly predicts 17 key points, cancelling the process from coarse to fine positioning, because the area of external key points is very large. If the third and fourth layers are added, a lot of time will be consumed. The final 68 key points of the face were superimposed by the output of two cascaded CNNS [15].

The main innovations of the algorithm are as follows:

1) The positioning of the face key point is divided into two parts of the internal key point and the external contour key point you to predict. This is very effective to avoid the problem of uneven feature point selection;

2) When detecting the internal points of the face, it does not use two cascaded CNNS for each key point to predict, as DCNN does, but only uses one CNN for each organ to predict, thus reducing the amount of calculation;

3) Compared with the traditional DCNN, it does not take the results of face collection as direct input, but adds a boundary localization for face border monitoring, which can effectively improve the network accuracy of coarse localization [9]. The structure and distribution area of the points are shown in Figure 3.

Figure 3. Point structure

The structure of the key points contains all the organs of the face respectively, including the contour of the eyes, nose, mouth and eyebrows. The green points in the figure are the key anchor points in the feature points. These key positioning points are the main forms of coarse precision positioning points, which are used to determine the rectangular position of the face, the distance between the outside of the eyes and the face, the distance between the two eyes, the distance between the eyes and the nose, the distance between the nose and the mouth, and the distance between the mouth and the chin. In addition, cheek feature points are included to distinguish the gender and race of a face. At present, the latest key points are 83 and 106 points, which are the embodiment of the convolution network with real fine accuracy. Because of the increase of key points, the facial information that can be recognized becomes richer and more in line with daily needs.

IV. Convolutional Neural Networks

Convolutional Neural Networks (CNNs/ConvNets) are very similar to normal neural networks in that they are composed of neurons with learnable weights and biases. The simulation diagram of common neural network is shown in below Figure 4.

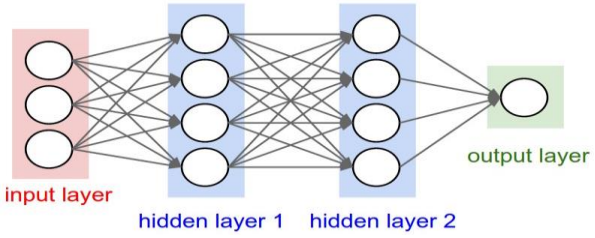


Figure 4. Traditional neural network diagram

Each neuron is involved in the calculation, and the dot product calculation is performed, and the output result is the score value of each class. The calculation method applicable in the traditional neural network is still used here. A convolutional neural network is composed of many layers, and the output is 3D. Some layers have parameters, and some layers have no parameters. General convolutional neural network contains the following layers.

Convolutional layer: Each layer in the convolutional neural network is composed of multiple convolutional units, and each parameter of the convolutional unit is optimized by the back propagation algorithm. Convolution operation is used to extract different input features. The first

layer can only extract some low-level features such as borders, lines and corners, while the multi-layer network is used to continuously screen and iterate complex features from low-level features. Pooling layer: usually, features with large dimensions will be obtained after the first layer is processed. The features will be cut into several areas to obtain the average or maximum value and new features with smaller dimensions will be obtained.

The reasonable Satisfaction Linear Units layer (ReLU layer) shall be provided to the reasonable Satisfaction Linear Units (ReLU) for nerve Activation function in this layer. Pooling layer: usually, features with large dimensions will be obtained after the first layer is processed. The features will be cut into several areas to obtain the average or maximum value and new features with smaller dimensions will be obtained. Fully Connected layer, which combines all local features to form a global feature, is used to calculate the final score of each type of feature. [10] The simulation diagram of convolutional neural network is shown in Figure 5 below:

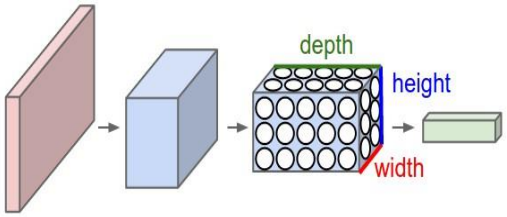


Figure 5. Convolutional neural network diagram

Modern mainstream machine learning libraries and interfaces, including TensorFlow, Keras, Thenao, Microsoft-CNTK, etc., can run convolutional neural network algorithms. In addition, some commercial numerical calculation software, such as MATLAB, also has the construction tool of convolutional neural network available.

V. Calling Platform API

The library of Python is used to send an HTTP request to the Face++ server, and upload a face picture captured by the camera. The server uses the feature point algorithm to analyze and process the picture, and returns the processed result as a object.

Figure 6 shows the execution flow of the program:

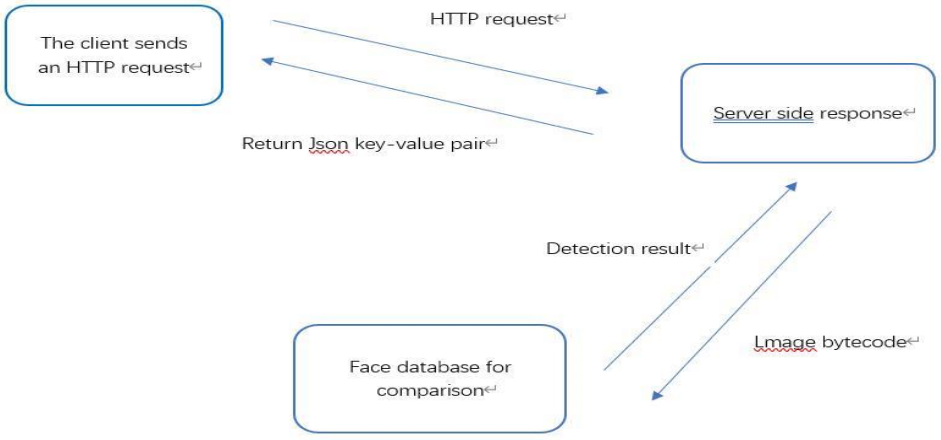


Figure 6. Flow chart of program execution

The program execution result is shown in below Figure 7.

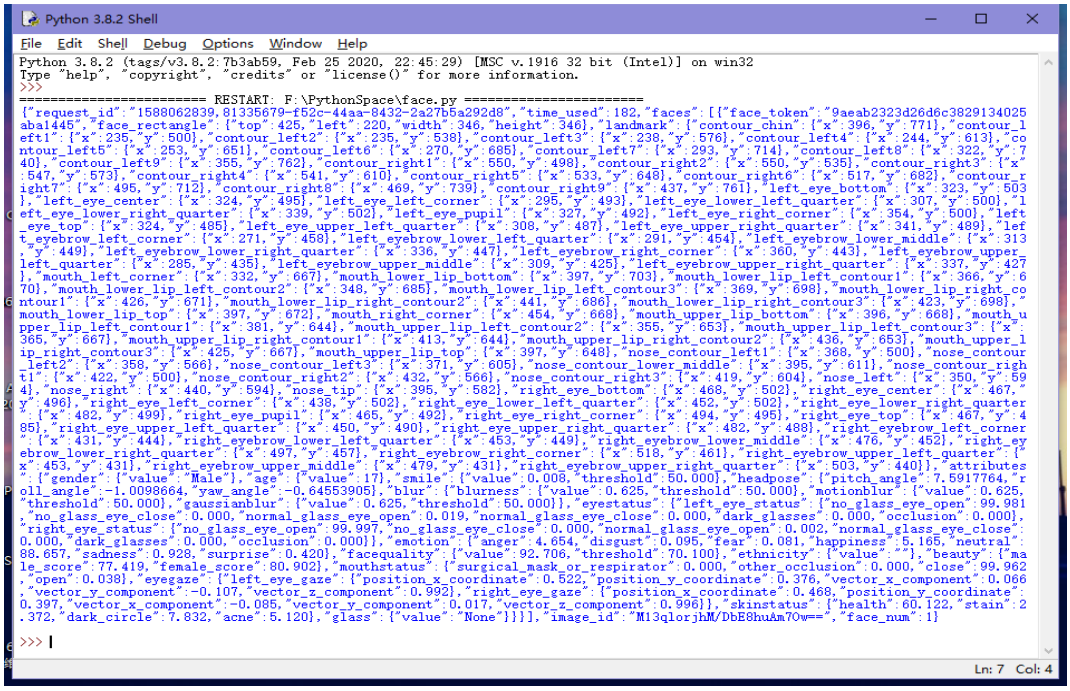


Figure 7. Program execution result diagram

The returned values are presented as Json key-value pairs. These return values contain information about the success or failure of the request, the cause of the request failure, and the error code information. The error code information contains both standard HTTP error code and API-specific error code information. When the request is sent successfully, the image information will be returned, when the size of the image exceeds the standard size, or the image path is wrong, the parameter abnormal information will be reported, when the image path and size are in line with the specification, continue to return the detected face information, there are several faces. The facial feature data in the return value includes facial clarity, facial posture, whether glasses are worn, light environment, and objective information such as age, race, gender, and appearance level scores.

Due to the different face data required under different environmental requirements, the face feature information returned by the API of Face++ platform contains many aspects, which can meet the general needs of the current market from different levels. Table 1 describes the returned values.

TABLE I. DESCRIPTION OF RETURNED VALUES

Field	Type	Instructions
Request_id	string	Distinguish each request, unique value
Faces	An array of	The array of faces that were detected
Image_id	string	The identification of the detected picture
Time_used	The integer	The time taken for the entire request, in milliseconds.
Error_message	string	Request failure information
Face_num	The integer	The number of faces recognized

The information in the returned faces array contains the face identifier, the position of the face rectangle box, the coordinates of the key points and the features of the face attributes, as shown in Table 2.

TABLE II. FACES ARRAY VALUES

Field	Type	Instructions
Face_token	string	Identification of human face
Face_rectangle	object	The position of the face box contains the following attributes. Top:The ordinate of the upper left corner of the face box left：The abscissa of the upper left corner of the face box width：The width of the face frame Height：The height of the face frame
Landmark	object	Coordinates of key points
Attributes	object	Facial feature information

After in solving the above errors, return to face feature information and standard test has great discrepancy in the document information, after the inquiry document to call API version and the latest version does not conform to, in the new version of the API, remove the glass this information, and will be integrated into the eye status information, this part of the information after the adjustment of API version, The correct facial feature information is shown below. After reviewing the technical documentation, the path to the server receiving the request was modified based on the updated version. However, the program still returns error information. After checking the error information, the initialization parameter error caused by coding problem was found. After comparing with the technical documentation of the developer platform, the coding format of the development environment was modified, and the request was successfully sent. After the request is sent successfully, the information returned is not the desired face feature information, and the API specific error code 413 is returned. After viewing the developer's technical documentation, it is found that when selecting the face image, the size of the image is not noticed, resulting in the image corresponding to the parameter cannot be correctly parsed.

Whether large Web system, Java App program, or a small Demo, have to undergo testing, only can be found in the test not appropriate, only when these problems will be improved, the program will be more introduction, strong, if too many redundant code, coupling relationship is too strong, will cause the program modification and optimization in the late very difficult.

Face recognition has many applications in people's production and life, initially as an alternative to fingerprint authentication. Compared with fingerprint authentication, face recognition authentication has the convenience of contact-free, which makes it easier for the public to accept [10].

But it's worth noting that when we authorize an app to identify us, it's likely to bundle our facial information with private identity information or even address information. This often means that when our face information is leaked, other bound information will be leaked along with the face information. An American company specializing in AI faces processing and recognition technology, has designed and released a face recognition application Clearview AI as shown in below Figure 8.

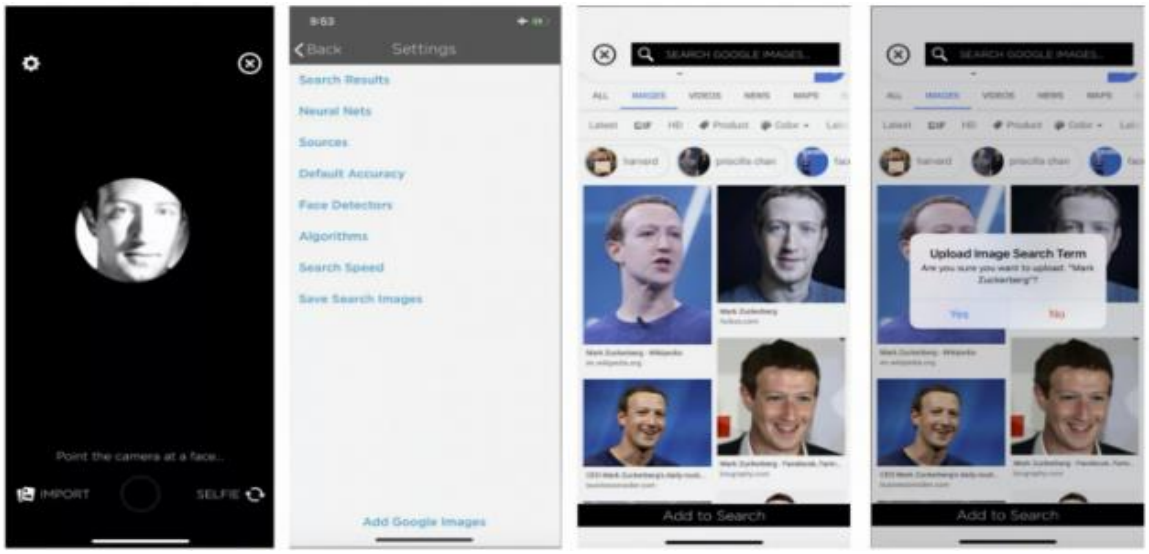


Figure 8. Facial recognition social program

Depending on the software to obtain the user's private pictures, and then upload them, you can normally browse the person's photos, as well as point to the specific information and shooting point of these photos. This photo-based technology isn't big news. Back in 2016, FindFace, a Russian startup, enabled matching a photo with a profile photo on Vkontakte, a Russian social network similar to Facebook. I can find strangers' social media accounts [13]. Although the App claims to help with "networking and intentional building". But the problem is that this is a huge database of faces. If Find Face fails to properly use the data -- school profiles, corporate customer lists, building comings and goings, mall customer movements, and so on -- harassment messages, death threats, and other incidents may occur, and facial recognition technology may violate our privacy in ways we can't imagine.

We all know that in the application of face detection identity, the information of a person's face is like a password, the password can be cracked, of course, the face can also be. As face recognition technology is widely used in applications such as face scanning authentication and face scanning payment, news of various face recognition systems being cracked continues to emerge. 3D structured light in vivo detection for the iPhone X app has been described as the most secure facial recognition detection technology in use today, but in 2017, a tech company managed to trick Apple's Face ID into using a human face mask produced by a 3D printer, as shown in Figure 9 below [10][11].

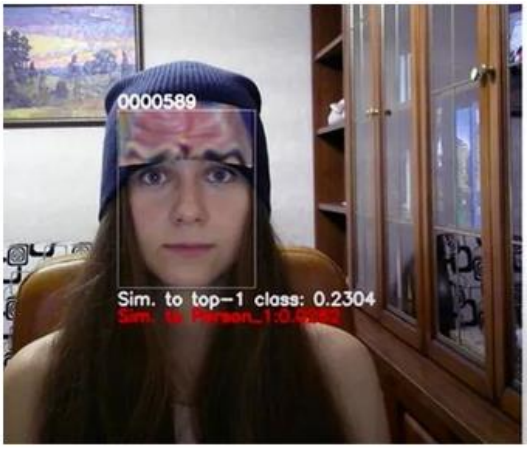


Figure 9. "Patch" cracking Face ID

Due to the technology cost is too high, we can now access to the most widely used facial recognition program is not structured light 3 d live

detection, and is only a 2 d plane detection or motion capture recognition, so we can often hear such as abundant nest "brush face take" with an inch of photo identification by criminals, intelligent security doors can use synthetic video unlock success. Considering the cost of cracking methods, the cracking methods of 3D printing materials are more flexible, but the experimental cost is too high. From the perspective of the success rate of recognition, planar photos cannot crack recognition.

Recently, a very clever method popular in machine learning technology circles is to confuse the recognition system by adding redundant image information that is difficult to recognize by the naked eye. This method is a way to trick the recognition operating system into producing errors. In 2019, researchers at Huawei's Moscow research center used a printer to print a partial image of a face and then used it to identify the face in a recognition system, an attempt that tripped up many well-known face recognition systems.

Future application prospects and a large number of benefits let face recognition technology is constantly touted by the capital market, but also invisible to promote the exploration of technology enterprises for emerging technologies. In order to promote their own face recognition system, many enterprises develop a new way, constantly for the development of face recognition technology of multiple application scenarios.

But as we all know, the current face recognition technology still has a lot of defects, the effect of many face recognition systems is not very good, which leads to the face recognition system in some scenes that do not need to use it is abused. The wrong perception that the main app does not see the effect also exposes the risks of such blind use. Such as some schools use face recognition door card machine, its starting point is for the sake of students' safety, but the technology adopted by the belong to the false face recognition technology, is not calculated according to the feature points or vector template, but simple through photo database to compare, so not only do not implement the initial security purposes, but the way of personal information may be leaked students, There are many non-compliant face recognition machines on the market. As shown in Figure 10.

Figure 10. Campus access control system

VI. Conclusion

Although face recognition technology brings a lot of convenience to our life, but also let us have a broader imagination for the future, but in the history of the battle between machine and human, most people win, so we need to ensure that the new technology is more accurate application, rather than no standard of active attempt. Due to the special circumstances of the pandemic this year, masks have become a necessity and are required to be worn in any public place. However, it also brings some troubles for people to be sentenced, wearing masks, mobile phones can not be unlocked, station access can not be identified.

Earlier, Danielle Baskin, a product designer, released a mask that looks like a Face patch, allowing you to unlock your phone with Face ID. In short, patches are used to complete incomplete facial information so that the machine can recognize it.

Although, as Danielle Baskin admitted on Twitter, the recognition rate is not guaranteed at this stage. However, once the news was released, many people doubted the reliability of face recognition technology. If someone wore a mask with their face information to unlock their phone, could it be unlocked? Therefore, for the future of face recognition, we should be in a state of experimentation, can not blindly expand the scope of use, if wrong, the impact may be world-class.

References

[1] Rui Gu. Overview of Occlusion Face Recognition Technology, 2018.

[2] Li-jun Yu. Application of Face Recognition Technology in the Exam Identity Authentication System, 2017.

[3] T.F. Cootes, C.J. Taylor, D.H. Cooper, et al. Active Shape Models-Their Training and Application [J]. Computer Vision and Image Understanding, 1995, 61(1):38-59.

[4] G. J. Edwards, T. F. Cootes, C. J. Taylor. Face recognition using active appearance models [J]. Computer Vision — Eccv', 1998, 1407(6):581-595.

[5] Cootes T F, Edwards G J, Taylor C J. Active appearance models[C]// European Conference on Computer Vision. Springer Berlin Heidelberg, 1998:484-498.

[6] Dollár P, Welinder P, Perona P. Cascaded pose regression [J]. IEEE, 2010, 238(6):1078-1085.

[7] Sun Y, Wang X, Tang X. Deep Convolutional Network Cascade for Facial Point Detection[C]// Computer Vision and Pattern Recognition. IEEE, 2013:3476-3483.

[8] Zhou E, Fan H, Cao Z, et al. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade[C]// IEEE International Conference on Computer Vision Workshops. IEEE, 2014:386-391.

[9] Zhang Z, Luo P, Chen C L, et al. Facial Landmark Detection by Deep Multi-task Learning[C]// European Conference on Computer Vision. 2014:94-108.

[10] Wu Y, Hassner T. Facial Landmark Detection with Tweaked Convolutional Neural Networks [J]. Computer Science, 2015.

[11] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks [J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.

Score Level Fusion for Iris and Periocular Biometrics Recogniton Based on Deep Learning

Yufei Wang
School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shaanxi 710032, Chain
E-mail: 649728545@qq.com

Bo Liu
School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shaanxi 710032, Chain
E-mail: 2502341201@qq.com

Songze Lei
School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shaanxi 710032, Chain
E-mail: lei-sz@xatu.edu.cn

Huan Zuo
School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shaanxi 710032, Chain
E-mail: 1094282550@qq.com

Yonggang Li
School of Computer Science and Engineering
Xi'an University of Technology
Xi'an, Shaanxi 710032, Chain
E-mail: 1984130837@qq.com

***Abstract*—Traditional iris recognition has high recognition accuracy and low misrecognition rate. However, in the case of mobile terminal or distance, the image resolution and image quality decrease, and the recognition rate also decreases. To solve the above problems, this article is based on deep learning technology, on the basis of single mode state recognition, from different levels of multimodal integration, the iris and the eyes in the score level fusion recognition research, put forward the adaptive dynamic weighted score fusion method, to determine the weighing values can adaptive algorithm of the modal, without artificial specified, dynamic weighting algorithm more flexible, stronger applicability. Experimental results of casIA-Iris-LAMP and CasIA-Iris-Distance Iris database in Chinese Academy of Sciences show that the proposed fusion algorithm has higher recognition accuracy and better recognition performance than the single mode recognition algorithm and the traditional fractional fusion method, which proves the effectiveness of the algorithm.**

Keywords-Multimodal Fusion; Iris Recognition; Periocular Recognition; Score Level Fusion

I. INTRODUCTION

With the development of the Internet, human identification technology has also become more

important. Traditional identification methods such as passwords and ID cards have many defects and cannot meet people's needs for security. Biometric-based identification technology uses the inherent physiological characteristics of the human body, such as: iris, face, fingerprint, voice print, gait, etc., to automatically identify a person through a computer, which is more secure and convenient, with the advantages of not forgetting, not losing and irreplaceable. As a biometric feature, the iris has the advantages of high uniqueness, stability and non-intrusiveness.

However, in some specific situations, such as mobile, long distances and weak constraints (where the person is moving slowly), it is difficult to capture a complete and clear iris texture, which will lead to reduced recognition rates. To solve this problem, the periocular area can be fused with the iris for multimodal fusion recognition. The periocular area, as a local feature of the face, is perfectly integrated with the iris position. When the iris image is acquired using a mobile device, it may not be possible to acquire the whole face, whereas the periocular region image can be acquired simultaneously with the iris, making the periocular region a better choice compared to the face. Therefore it is important to study the fusion of iris and periocular recognition methods to improve recognition accuracy and stability and to overcome the anti-forgery nature of unimodal feature recognition techniques.

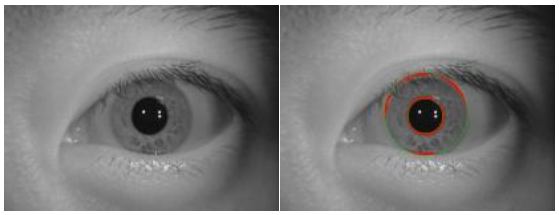
Liu et al [1] were the first to use convolutional neural networks for iris recognition and obtained more superior recognition results. mohtashim Baqar et al used an improved deep belief network for iris recognition [2]. Zhao et al used a capsule network in the network structure and improved the network structure for iris feature extraction and recognition to improve the problem of small iris samples [3] Zhao and Kumar et al [4] proposed a new network model that recovered periocular features by combining semantic information, where higher matching accuracy was achieved using a small number of samples and better recognition results were obtained. proenca and Neves et al [5] improved the performance of periocular recognition based on discarding iris and sclera features in visible light and recognising other regions around the eye. Performance of periocular recognition. Chen et al [6] used additive rules to fuse two biometric features, face and iris, in the score level for fusion recognition, and used a wavelet neural network classifier for classification recognition. The experimental results demonstrated that multimodal biometric verification is more reliable and accurate than single biometric verification. Son et al [7]fused two biometric features, face and iris, in the feature layer Direct linear discriminant analysis was used in the fusion to reduce the dimensionality of the feature information, and finally nearest neighbours were used for classification and recognition. Teddy Ko et al [8] discussed the impact of multiple fusion methods and image quality on the overall recognition accuracy on fingerprints, faces, and irises. Zhang et al [9] designed a deep feature fusion network using a Maxout structure for feature extraction that The weighted fusion on the feature layer for iris and face achieved good recognition results, and the model occupies less storage space and is suitable for mobile.

Therefore, this paper proposes an adaptive weighted fusion method based on iris and periocular score layers based on deep learning techniques. It has been experimentally demonstrated that the fusion method can significantly improve the recognition accuracy compared with unimodal recognition.

II. FUSION USION METHODS

A. Iris and periocular image pre-processing

Iris image preprocessing mainly includes two steps: segmentation and normalisation. In this paper, the iris is segmented using the Osiris V4.1 system [10]. Osiris uses the Viterbi algorithm [11], which is able to perform at low resolution, and which is able to quickly locate the iris contour. The green circle indicates the outline of the iris and the red marker is the non-iris portion of the annular region. This is shown in Fig.1.



(a)Original image (b)Marked images

Figer 1. Iris pre-processing diagram

(a)Iris Chart (b)Normalised iris image

Figer 2. Iris normalisation

The iris normalisation process is carried out with the aid of a rubber sheet model, which is then used to obtain a rectangular iris image for classification recognition. As shown in Fig.2, where (a) shows the original image of the iris after segmentation, as a circular region, and (b) shows the normalised iris image, after which the iris is converted to a rectangular image.

Pre-processing of the periocular area includes periocular angle correction and periocular normalisation. The periocular image is downsampled from 640 ×480 pixels to 224 ×224 pixels by normalisation. In order to prevent the periocular image from not being wide or tall enough during normalisation, this paper uses bi-cubic interpolation, using the surrounding skin texture to fill in. A normalised periocular image with higher feature differentiation is obtained.

B. Traditional score level fusion methods

Traditional score layer fusion methods include those based on combined strategies [12][13] and manual weighting methods. The commonly used combinatorial decision strategy fusion methods are: maximum value, minimum value, multiplication rule, and addition rule.

1) Maximum value rule

$$f_i = \max \ n_i^m \tag{1}$$

2) Minimum value rule

$$f_i = \min \ n_i^m \tag{2}$$

3) Multiplication rules

$$f_i = \prod_{m=1}^{M} n_i^m \tag{3}$$

4) Addition rule

$$f_i = \sum_{m=1}^{M} n_i^m \tag{4}$$

In the above equation n_i^m denotes the matching score of the mth modality after normalization. f_i denotes the score after fusion.

5) Manual setting of weights

In this paper, before fusing the two modalities of iris and periocular, the weights of both are divided and the iris and periocular weights are set to weighted fusion with weighting ratios from 1:9 to 9:1 to obtain the fusion accuracy under each weighting ratio. The weighting formula is shown in Equation (5).

$$F_{score} = \alpha I_{score} + \beta E_{score} \quad (5)$$

F_{score} is the fused score, I_{score} denotes the iris matching score and E_{score} denotes the periocular matching score. α and β are the weighting coefficients for the iris and periocular modal scores respectively, taking values in the range [0,1], and α+β=1, set by hand. The method was used to find the best matching weighting coefficients through multiple experiments.

C. Our iris and periocular score level fusion

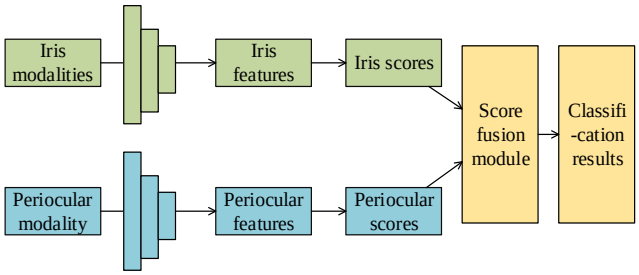
Score layer fusion is a method to fuse the scores of different biometric images or data after feature extraction and feature matching. The fused score is then used for identification. The main idea of score layer fusion is to use combinatorial decision strategy, which includes addition, multiplication, weighted sum, Max and min rules, etc.. Compared with the feature layer fusion, the score layer fusion does not damage the integrity of each biometric recognition, and is easier to carry out the fusion. The fusion method is simple and the fusion speed is fast. However, the score layer fusion will lose a lot of feature information, and it is difficult to achieve higher recognition effect.

Biological characteristics themselves have very rich characteristic information, and there are obvious differences in the characteristics of different organisms. In single mode recognition, due to the limitation of acquisition equipment, the surrounding environment, acquisition distance and other influences, the collected biometric image data received serious interference, resulting in a lot of adverse impression for the final recognition. Therefore, multiple modal biometrics can be combined to enhance the diversity of feature information and improve the recognition performance. According to the level of fusion, pixel layer fusion contains the most abundant feature information. Due to the complexity and diversity of data types, it is difficult to operate during the fusion, and data cannot be effectively fused. Feature layer fusion is to fuse the extracted feature data together on the basis of the extracted feature data. Therefore, some features of the fused data are lost. Score layer fusion uses the score obtained after feature matching, and feature information loss is large. However, the score data is easy to obtain and the fusion operation is simple. It has become the most widely used multi-mode fusion technology in deep learning. One difficulty of score layer fusion is that too many distinguishing features are lost in the fusion process and too little feature information can be used, which limits the fusion performance. In order to make up for the loss of characteristic information on the score layer as much as possible. In this chapter, when selecting score layer fusion, more feature information is extracted through deep learning, and the fusion strategy is improved by adding weights to improve the performance of score layer fusion recognition.

Score layer fusion is the most commonly used method in multimodal fusion. The fusion makes full use of the score information of different modalities and does not result in dimensional disasters. In this paper, we propose an algorithm for fusion of iris and periocular score layers based on a combined decision strategy with set

weighting factors. The fusion problem of the iris and periocular modalities is regarded as a combined score determination problem, and the fusion rules are analysed in depth. By conducting theoretical and experimental studies on the information of each modality during fusion and the fused scores, a fusion formula that can automatically obtain the weight coefficients of each modality is finally obtained.

Fig.3 shows a schematic diagram of iris and periocular fusion in the score layer. First, the iris is extracted by performing iris segmentation on the acquired periocular image. Image pre-processing is performed on the iris and periocular images to obtain the two modalities of iris and periocular that can be used for recognition. Next, the deep learning algorithm is used to extract and match the features of the iris and periocular modalities to obtain the iris and periocular scores respectively. Finally, the scores of the two modalities are fused and recognised by the score fusion module to obtain the final classification result.



Figer 3. Schematic diagram of scorer layer fusion

Aiming at the problems exposed by the traditional combinatorial decision fusion strategy, a weighted score layer fusion algorithm is proposed in this section. The algorithm fully considers the importance of the matching score of each mode to the recognition result. By dynamically adding a weight coefficient to the fraction of each mode, the fusion recognition is carried out at last. The algorithm studies the importance of each modal fraction to the recognition result, which improves the recognition accuracy significantly, alleviates the over-fitting problem, and improves the overall recognition performance.

The most widely used existing method is to use manually set filters to obtain features and finally a distance metric to obtain a matching score. Commonly used distance similarity methods are: cosine distance, Hamming distance and Euclidean distance. In this paper, we start from deep learning and feature extraction of iris and periocular images by building deep neural networks. In this paper, ResNet18 is used as the feature extraction network with Softmax function and Categorical_crossentropy loss function to achieve the matching of iris and periocular modal scores.

In the last layer in the neural network, this paper uses the Softmax function to implement the function of neural network classification. In a multi-classification network, Softmax implements the function of normalisation and the output is a vector of predicted probabilities, indicating the probability of the input data belonging to each category, its beneficial to the data fusion study in this paper. The Softmax formula is expressed as shown in equation (6).

$$y_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{6}$$

In the above equation, z denotes the input data of the last layer and z_i denotes the ith data in z. After Softmax calculation on the input data, the obtained data is normalised to between [0,1], which is the probability value of each class.

The recognition study in this paper belongs to

the domain of multi classification, so Categorical_crossentropy is used as the loss function of the fusion network to calculate the loss of real and predicted data, and its expression is shown in Equation (7).

$$L = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i=0}^{N-1}\sum_{C=0}^{C-1} y_{ic}\log p_{ic} \quad (7)$$

In the above equation, N denotes the number of overall samples and K denotes the number of final classifications. The y_{ic} is a symbolic function taking the value 0 or 1. The meaning of p_{ic} is the predicted probability c of the first category.

In this paper, score layer fusion is performed in a weighted manner. The iris and periocular are fused on the score layer by an automatic weight setting method. By setting the weights at the time of fusion, it is hoped to obtain better recognition results than under unimodal recognition and simple rule fusion.

On the acquisition of adaptive weights. In this paper, adaptive weighting is applied to the iris and periocular scores by analysing the correlation between each modality and ultimately the degree of importance to the recognition result. The formula for obtaining the weights for the iris is shown in Equation (8).

$$I_{weight} = \frac{I_{score}}{I_{score} + E_{score}} \quad (8)$$

The formula for obtaining the periocular weights is shown in Equation (9).

$$E_{weight} = \frac{E_{score}}{I_{score} + E_{score}} \quad (9)$$

The overall fusion formula is shown in Equation (10).

$$F_{score} = I_{weight} \times I_{score} + E_{weight} \times E_{score} \quad (10)$$

F_{score} is the fused score, I_{score} denotes the iris matching score and I_{score} denotes the periocular matching score. The weights are set by adaptive weighting so that each modality plays a maximum role in the recognition result when fused. Thus, the recognition effect is improved.

III. Experimental Results and Analysis

In this paper, the CAS database is selected for the exploration of score layer fusion algorithm. The experimental results of periocular and iris recognition under unimodality are shown first, followed by the fusion experimental results of additive rule, maximum value rule and minimum value rule respectively, and finally the experimental results after score layer weighting are shown.

A. Experimental protocol

The experimental hardware system version of this paper is Linux Ubuntu16.04. A Navidia GeForce RTX2080Ti GPU is used for accelerated computing to speed up the convergence of the network model and increase the computing speed. In addition, version 3.6.8 of python environment was chosen. The Tensorflow framework was used for network construction. The specific configuration is shown in Table 1. This table mainly describes the configuration of the processor, graphics card, operating system,

framework and compiler.

TABLE I. ENVIRONMENT CONFIGURATION

Hardware Environment	Processor	Inter(R) XeonGold b6254
	Graphics card	Navidia GeForce RTX2080Ti
Software Environment	Operating System	Linux Ubuntu 16.04
	Framework	Tensorflow-GPU
	Programming language	Python3.6.8
	Compiler	PyCharm

In this study, two Iris datasets from Chinese Academy of Sciences were selected: CasIA-Iris-LAMP and CasIA-Iris-Distance. The CasIA-Iris-LAMP dataset includes 791 categories of Iris images, with a total of 15,820 images. In these Iris images, there are also an equal number of corresponding periocular images. Casia-iris-distance data set includes 284 categories of Iris images, with a total of 5073 images. Among these Iris images, an equal number of corresponding periocular images are also included. The above data set is divided in a 3:1:1 ratio. The Casia-Iris-LAMP dataset contains 791 categories, including 9492 training sets, 3164 validation sets and 3164 test sets respectively. Casia-iris-distance data set contains 284 categories, including 3160 training sets, 1024 verification sets and 889 test sets. The image size is 224 pixels ×224 pixels. Detailed parameters are shown in Table 2.

TABLE II. EXPERIMENTAL DATA SETS

Dataset	Category	Iris images	Periocular images	Training set	Validation set	Test set
CASIA-Iris-Lamp	791	15820	15820	9492	3164	3164
CASIA-Iris-Distance	284	5073	5073	3160	1024	889

In this paper, score layer fusion is used. After obtaining pre-processed iris image and pre-processed eye image, model parameters need to be set in advance. Before fusion, iris and periocular recognition models in single mode were obtained, and the model training parameters were shown in Table 3. The input dimension is 224×224×3, the number of images input by the model is 32, the number of steps trained by the model is 2000, and the gradient descent function used by the model is RMSprop function.

TABLE III. EXPERIMENTAL CONFIGURATION

Relevant parameters	Parameter settings
input_shape	224×224×3
batch_size	32
epoch	2000
optimizers	rmsprop
learning_rate	10^{-3}
rho	0.9
decay	10-6

Before the experiment, we selected the iris data set of Chinese Academy of Sciences,

preprocessed the iris image and periocular image in advance, and configured the experiment in detail. In the performance test of score layer fusion, this paper mainly uses the method of calculation accuracy to evaluate the accuracy of single mode algorithm and score layer fusion algorithm.

B. Results of fusion experiments

The experiments in this section include unimodal periocular and unimodal iris recognition, as well as multimodal score layer fusion. The fusion of score layers compares the traditional methods based on the maximum value, minimum value, multiplication rules, addition rules in the simple combination strategy and the adaptive weighted rules proposed in this paper, respectively, and also compares the results of iris and periocular recognition in unimodality. The experimental results are shown in Table 4.

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS

Dataset	Fusion rule	Recognition accuracy
CASIA-Iris-Lamp	Maximum value	0.992055
	Minimum value	0.955830
	Multiplication	0.992017
	Additive	0.992373
	Unimodal periocular	0.982617
	Unimodal iris	0.981252
	Adaptive weighting	0.993831
CASIA-Iris-Distance	Maximum value	0.961754
	Minimum value	0.866141
	Multiplication	0.972512
	Additive	0.961754
	Unimodal periocular	0.933633
	Unimodal iris	0.894263
	Adaptive weighting	0.985376

Several conclusions can be drawn from the experiments in Table 4, as follows.

1) In terms of the dataset, the recognition accuracy on the CASIA-Iris-Lamp dataset is higher than that on the CASIA-Iris-Distance dataset, regardless of whether the recognition is performed under multimodal recognition or unimodal recognition. Moreover, the recognition effect of unimodal periocular is better than that of unimodal iris recognition.

2) In terms of modality, the recognition accuracy of score layer fusion was above 98.9% on both CASIA-Iris-Lamp dataset and 98.2% on both unimodal recognition accuracies. On the CASIA-Iris-Distance dataset, the recognition accuracy of score layer fusion was above 97% in all cases, and the unimodal recognition accuracy was lower. Therefore, it can be concluded that the recognition accuracy of periocular and iris score layer fusion is higher and better than that of unimodal periocular and unimodal iris recognition, except for the minimum value fusion rule.

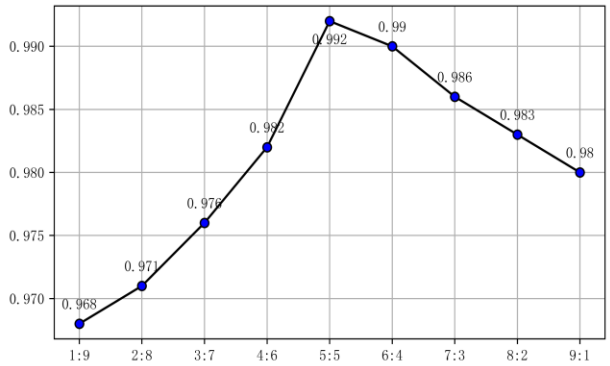
3) In terms of fusion rules, it was shown by experimental results on both datasets. The traditional score layer fusion has the best recognition by the multiplicative rule and the worst recognition by the minimum value rule. The adaptive weighted score layer fusion proposed in this paper achieves the best recognition results.

In this paper, we set the weight range to [0.1, 0.9] in the manual weighted fusion experiments for iris and periocular in the score layer. The weights are set and calculated in order from 1:9 to 9:1 for the iris and periocular. The weights were first set to 0.1 for the iris and 0.9 (1-0.1) for the periocular area, and were updated at a rate of 0.1 each time. The results of the experiments are shown in Table 5.

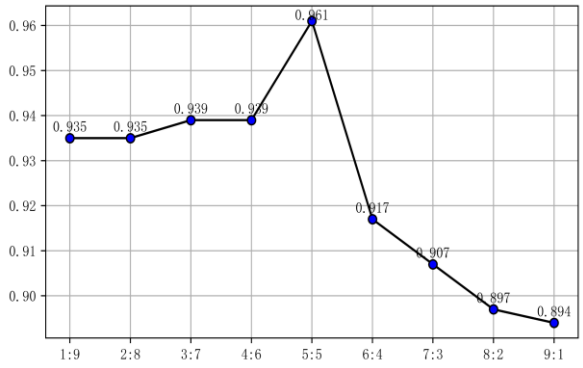
TABLE V. COMPARISON RESULTS OF DIFFERENT WEIGHTS

DataSet	CASIA-Iris-Lamp	CASIA-Iris-Distance
1:9	0.968541	0.935883
2:8	0.971083	0.935883
3:7	0.976167	0.939257
4:6	0.982840	0.9392575
5:5	0.992373	0.9617547
6:4	0.990149	0.917885
7:3	0.986653	0.907761
8:2	0.983158	0.897637
9:1	0.980934	0.894263

Table 5 combined with Fig.4 and Fig.5 shows that the fused recognition accuracy steadily increased as the weight of the iris increased until it reached an optimum of 99.2% at a weight of 0.5 for the iris. This experiment shows that on both the CASIA-Iris-Lamp and CASIA-Iris-Distance datasets, the best fusion scores were obtained for both iris and periocular modalities at a weighting ratio of 1:1. As the weight of the iris continued to increase, the fusion recognition rate gradually decreased to 98%. The experiments show that the recognition effect after fusion is strongly influenced by the iris. Good performance was also obtained with manually weighted fusion.



Figer 4. Fusion plot of different weights on CASIA-Iris-Lamp



Figer 5. Schematic representation of the fusion of different weights on CASIA-Iris-Distance

IV. CONCLUSIONS

In this paper, score layer fusion of iris and periocular is studied from the perspective of multimodal fusion. In response to the degradation of recognition rate due to image quality degradation in mobile or remote situations, a score layer fusion method based on adaptive weighting was proposed, where the weights of the modalities can be determined adaptively. Experimental comparisons of the methods were carried out in the CAS iris library. The experimental results show that the adaptive weighted score layer fusion achieves better recognition results, demonstrating the effectiveness of the algorithm in this paper, and also proving that the accuracy of the score layer-based multimodal fusion algorithm is significantly better than that of the single-modal recognition algorithm.

ACKNOWLEDGEMENT

This work was supported by National Joint Engineering Laboratory of New Network and Detection Foundation (grant no. GSYSJ2016008). The experiments in this paper use the CASIA-Iris collected by the Chinese Academy of Sciences' Institute of Automation (CASIA).

REFERENCES

[1] Liu N , Zhang M , Li H , et al. DeepIris: Learni-ng Pairwise Filter Bank for Heterogeneous Iris Ver-ification [J]. Pattern Recognition Letters, 2015, 82(2):154-161.

[2] Baqar M, Ghani A , Aftab A , et al. Deep belief networks for iris recognition based on contour detection[C]// International Conference on Open Source Systems & Technologies. IEEE, 2017.

[3] T. Zhao, Y. Liu, G. Huo and X. Zhu, "A Deep Learning Iris Recognition Method Based on Capsul-e Network Architecture," in IEEE Access, vol. 7, pp. 49691-49701, 2019, doi: 10.1109/ACCESS.2019.2911056.

[4] Zijing Zhao,Ajay Kumar 0001. Accurate Periocular Recognition Under Less Constrained Environment Using Semantics-Assisted Convolutional Neural Network. [J]. IEEE Trans. Information Forensics and Security, 2017, 12(5).

[5] Hugo Proença,João C. Neves. Deep-PRWIS: Perioc-ular Recognition Without the Iris and Sclera Using Deep Learning Frameworks. [J]. IEEE Trans. Infor-mation Forensics and Security, 2018, 13(4).

[6] Chen C H, Chu C T. Fusion of Face and Iris Features for Multimodal Biometrics[C]// Internation-al Conference on Advances in Biometrics. Springer-Verlag, 2006.

[7] Son B, Lee Y . Biometric Authentication System Using Reduced Joint Feature Vector of Iris and Face[C]//International Conference on Audio-&Video-based Biometric Person Authentication. Springer-Ver-lag, 2005.

[8] Ko T .Multimodal biometric identification for large user population using fingerprint, face and iris recognition[C]// Applied Imagery & Pattern Recogn-ition Workshop. IEEE, 2005.

[9] Qi, Zhang, Haiqing, et al. Deep Feature Fusion for Iris and Periocular Biometrics on Mobile Devices[J]. IEEE Transactions on Information Forensics & Security, 2018.

[10] Othman N, Dorizzi B, Garcia-Salicetti S.OSIRIS: An open source iris recognition software [J]. Pattern Recognition Letters, 2016, 82(oct.15):124-1

[11] Sutra G, Garcia-Salicetti S, Dorizzi B. The Viterbi algorithm at different resolutions for enhanc-ed iris segmentation[C]// Iapr International Confere-nce on Biometrics. IEEE, 2012.

[12] Snelick Robert,Uludag Umut,Mink Alan,Indovina Michael,Jain Anil. Large-scale evaluation of multim-odal biometric authentication using state-of-the-art systems. [J]. IEEE transactions on pattern analysis and machine intelligence, 2005, 27(3).

[13] Anil Jain,Karthik Nandakumar,Arun Ross. Score normalization in multimodal biometric systems [J]. Pattern Recognition, 2005, 38(12).

Research on Static Gesture Recognition Based on Deep Learning

Min Zhang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1565364293@qq.com

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

***Abstract*—With the continuous development and progress of the times, the ways of human-computer interaction have become more and more diverse. In order to reduce the spread of the new crown virus, gesture recognition has become a hot topic in the field of human-computer interaction in recent years. Traditional gesture recognition is affected by the environment and database, etc., with poor robustness and low accuracy. In order to improve the recognition rate of static gestures, this paper proposes to establish a deep learning model using CNN convolutional neural network, and a static gesture recognition method based on template matching. By establishing a palm template diagram, the gesture image to be recognized is matched with the template diagram based on the feature point, and the image is rotated after matching, and the template based on the grayscale value is matched again, so as to extract the gesture part. Through experimental proof, the algorithm can effectively improve the gesture recognition rate, the recognition accuracy rate reached 93.17%, and the recognition speed is faster.**

Keywords-Static Gesture Recognition; Template Matching; Deep Learning; Convolutional Neural Networks

I. INTRODUCTION

Gesture recognition, as the name suggests, refers to the recognition of the palm and arm parts of the human body and the reading of the meaning expressed by the gesture. Gestures are generally divided into two forms: dynamic and static [1]. Dynamic gesture refers to the trajectory of palm movement, this one trajectory change represents a combination of gestures, for the recognition of combination gestures, that is, dynamic gesture recognition, common dynamic gesture recognition acts on: control the synchronous movement of the robot arm and the human gesture, control the page jump through the gesture movement, etc. Static gesture refers to a single gesture, the recognition of a single gesture can read continuous information, common static gesture recognition acts to assist deaf people and ordinary people to communicate, through gesture reading for simple verification. Static gestures are an extremely convenient human feature with the advantage of being simple and easy to read.

At present, the commonly used gesture acquisition methods are: reading gestures through data gloves [2], but the data gloves collect data, and the final recognition rate is easily affected by the device. Reading gestures by optical marking, which transmits the position of the human hand

and the dynamic changes of the fingers to the system screen through infrared, also achieves good gesture recognition, but still requires complex devices to support the intervention of external devices. The Bayesian attention model [3] is used for gesture detection and then the support vector machine (SVM) is used to complete gesture recognition, which is too complex and less practical.

Based on the gesture recognition method of the hidden Markov model [4], this method has achieved good results, but the real-time and accuracy rate need to be improved. Deep learning [5][6] has made great achievements in the field of computing. In neural networks, large network frameworks such as R-CNN [7], Faster R-CNN[8], YOLO[9], and SSD[10] can be well used to recognize gestures, but the amount of computation in large networks is too large to use gesture recognition. Therefore, gesture recognition based on computer vision has become the mainstream research direction. This experiment will capture the gesture image through the camera, realize the processing of the gesture image in the computer, segment the gesture profile, extract the gesture part, and then take the method of deep learning, establish a deep learning model and identify it, and realize the effect that the computer can "recognize" the gesture.

II. ESTURE RECOGNITION ALGORITHM

A. Gesture image acquisition

The gesture image is captured through the camera, and the gesture image is collected into a computer, that is, the gesture image is converted into a matrix of pixels stored in the computer, and pixels are a matrix composed of numbers. In a computer, the operation of an image is, in essence, an operation of a matrix of pixels.

B. Overall system framework

As shown in Fig. 1, the overall framework of the system is divided into four parts: image preprocessing, gesture segmentation, gesture feature extraction and hand recognition. The gesture image is captured through the camera, and the color image is passed into the computer, which requires certain processing of the image to make the overall image easier to extract the gesture. The main role of gesture segmentation is to separate the hand area from the background, and a large part of the performance of this part affects the main performance of the system; Gesture feature extraction is to reduce the depth of the recognition network of hand extraction contour features, reducing the training time; The hand recognition network identifies what kind of gesture is from manually extracted features.

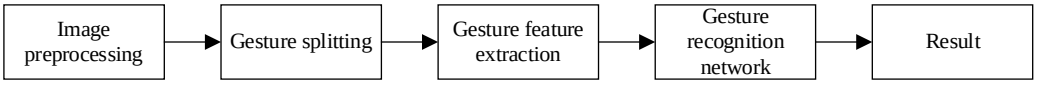


Figure 1. System framework

C. Gesture image preprocessing

The gesture image is captured through the camera, and the color image is transmitted to the computer, and the image needs to be processed to make the overall image easier to extract the gesture, and then extract the gesture part of the image through a certain algorithm, and then judge the extracted gesture to complete the recognition of the gesture.

1) Gesture image grayscale

Color images in the computer, generally stored in the form of RGB three-color channels, grayscale of the image, refers to the image changed to a single channel form stored in the

computer, can effectively reduce the amount of computing during image processing, improve processing efficiency [11].

Grayscale images have equal values in all three color channels of RGB. The grayscale method of the image used in this experiment is weighted averaging. The weighted averaging principle is as follows:

$$f(R,G,B) = Wr * R + Wg * G + Wb * B \quad (1)$$

$f(R,G,B)$ represents the grayscale, R, G, B, represent the value of the point on the three-color channel, Wr, Wg, Wb,respectively represent the weights of R, G, B.

2) Image enhancement processing of gesture images

Image enhancement refers to the specific processing of the image to make the image suitable for a specific occasion. In this experiment, the static gesture recognition needs to make the features of the gestures in the image more prominent, so as to improve the recognition efficiency and effect of the gestures.

In the gesture image captured by the camera, there will be many noise points that interfere with the recognition of the image, and the image smoothing processing is required, which can effectively reduce the interference of the noise points in the image and make the recognition of the image more accurate. The image smoothing method used in this experiment is: median filtering. Set the convolutional kernel with a size of 3×3, sort the 3×3 area centered on the pixel, and assign the median value to the pixel. It can be very effective in smoothing impulse noise while protecting the sharp corners of the image without affecting profile detection.

D. The gesture area splits the network

The hand area processing part is to manually extract features by dividing the hand area into hand areas. Because convolutional neural networks can avoid complex pre-processing of images, directly input the original image, and can identify changing patterns, tolerate image distortion, and have strong robustness [12]. The role of the convolutional layer of the gesture recognition part of the convolutional neural network is to extract the characteristics of the gesture in the picture data, if the picture is processed before the picture is sent to the neural network for discrimination, so that the feature extraction of the neural network requires less convolutional kernels, so as to improve the recognition rate while optimizing the network structure.

The flow of hand area feature extraction is shown in Fig. 2.

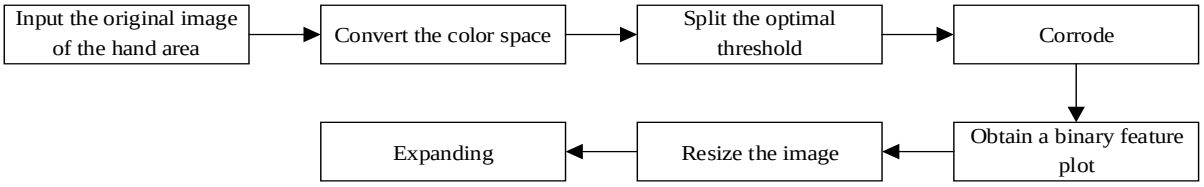


Figure 2. Feature extraction flowchart

The skin color pixels of the human body are in the Ycrcb space, the skin pixels basically meet the elliptical distribution, each pixel is analyzed, whether it is within the ellipse where the skin color pixels are located, if in the ellipse, it means that the pixel is the skin pixel and extracted. Where Y represents the brightness, Cr and Cb can be represented as the difference between the

blue component, the red component and the brightness, respectively, representing the chroma.

In this experiment, RGB pattern images are often converted to Ycrcb color space by the following formula and then processed:

$$Y = 0.299R + 0.587G + 0.114B \quad (2)$$

$$Cr = -0.147R - 0.289G + 0.463B \quad (3)$$

$$Cb = 0.615R - 0.515G - 0.100B \quad (4)$$

Then by adjusting the thresholds of Y, Cr, Cb, you can segment the gesture from the image.

E. Gesture feature extraction

The static gestures studied in this article have one thing in common – the palm of the hand. Only the palm part of the template matching, get the palm part in the image and the angle of deflection relative to the horizontal direction, distinguish the gesture area, only identify the gesture part, and then eliminate the impact of the non-gesture part (such as the face, etc.), the red line contains the area of the palm template and the gesture area respectively. For example: Fig. 3, Fig. 4.

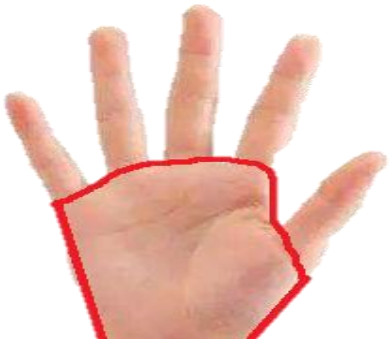


Figure 3. Palm area

Figure 4. Gesture area

This paper can match the palm area in the gesture image through SIFT features, calculate the palm area through the feature point coordinates, enlarge the palm area to a certain extent, increase a certain pixel in the horizontal and vertical directions, that is, include the entire gesture area, identify the gesture area, and effectively avoid the interference caused by the non-gesture part.

Skin tone areas in an image are extracted through skin color recognition to remove the influence of non-skin color parts. The sift algorithm then matches the template of the palm image and the input image, and the angle of deflection of the two images is calculated by the sift algorithm for the direction given by the feature point. Using the principle of image rotation, the pixel matrix is deflected to overcome the defects of the template matching algorithm based on grayscale values. Then the palm is matched with the image by the square difference matching method, the coordinates of the palm area are obtained, and the gesture area can be obtained by stretching the area. The flowchart of the template matching module is shown in Fig. 5.

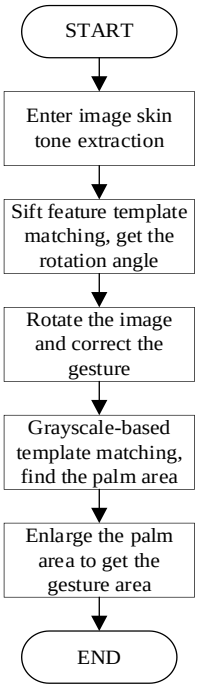


Figure 5. Template matching flowchart

The image rotation algorithm used in this experiment is to rotate the gesture image at a specified angle in the center. Gesture images are stored in a computer in the form of a matrix of pixels, so a rotated image is a matrix rotation algorithm. Therefore, if the width of the original figure is w, the height is h, and the rotation angle is θ, the following formula can be obtained:

$$x_1 = \frac{x_0 - w}{2}; y_1 = -y_0 + \frac{h}{2} \quad (5)$$

F. Gesture recognition network

Static gesture recognition neural network diagram as shown in Figure 6, through the input layer input gesture image, after each convolutional layer of the input gesture image convolution processing, connect a pooling layer, the input gesture image for local feature extraction work, after the processing of two layers of convolutional layer and two layers of pooling layer, through the processing of two layers of fully connected layers, the image features are classified. Through local connection and weight sharing, the construction of CNN convolutional neural network is realized.

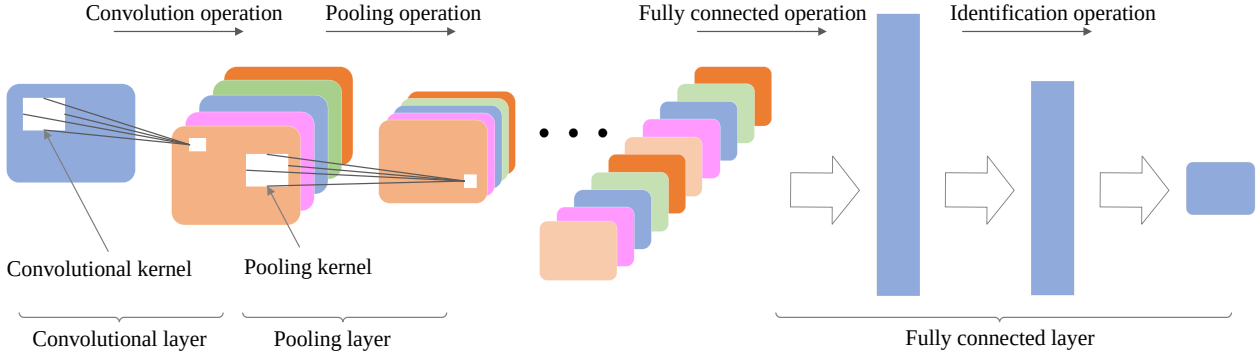


Figure 6. Static gesture recognition neural network framework

The construction of the convolutional layer is the core of the entire CNN construction [13], in the image recognition, set the size of the convolutional core, convolution operation on the input image, that is, the convolution core slides on all pixels on the image, and the pixel at that location is convoluted, and different features can be extracted for the image through the setting of the convolutional core. In the convolutional layer setting for static gesture recognition, the convolutional kernel is set to 2×2 size, the step

size is set to 1, and the mathematical principle of convolution is as follows:

$$h(x) = f(x) * g(x) = \int f(t) * g(x-t) d_t \quad (6)$$

Suppose a gesture image with 64 × 64 pixels, after convolutional and pooling, use the above equation to calculate, the output size is 32×32.

After the convolutional layer, the static gesture recognition model construction needs to add a pooling layer, the main role of pooling is to reduce the parameters and calculation times in the network, which can effectively improve the training efficiency of the static gesture training set and prevent over fitting during the training process.

In a static gesture recognition neural network, the relu function is used to make network training more efficient and increase the nonlinearity of the network compared to the sigmoid function.

The fully connected layer mainly implements the function of classification in the static gesture recognition neural network. The essence of full connection is to linearly transform the feature space acquired by the convolutional layer into another feature space, and connect the two layers of the full connection layer after the static gesture recognizes the convolutional layer in the neural network, and the weighting operation of the feature can effectively reduce the influence of feature location on classification.

III. THE TRAINING OF THE NEURAL NETWORK AND THE FINAL RESULT

A. Static gesture recognition preprocesses module test results

In order to verify the effectiveness of the recognition method proposed in this paper, this paper collects six kinds of gestures in life scenarios, namely digital gesture 1, digital gesture 2, digital gesture 3, digital gesture 4, digital gesture 5 and gesture good. After acquiring a certain number of gesture images through the camera, each picture is preprocessed to make the image easy to identify the gesture, and the size meets the input size of 89×89 pixels required by the deep learning model. After the image is processed, the effect is shown in Fig. 7.

Figure 7. Image preprocessing effect

Through skin color recognition, the skin tone of the human body is extracted from a complex background. Through the method of template matching, the coordinate position of the gesture in the image is found, and the non-gesture part is shielded, and then the extraction and pre-processing function of the gesture under the complex background is completed.

B. Static gestures identify deep learning model test results

In the training of the deep learning model of the static gesture recognition convolutional neural network, the recognition accuracy calculated by the model recognition verification set continues to increase with the increase of the number of trainings, and the final recognition accuracy is stable between 96.9% and 100%. The effect during training is shown in Fig. 8, Fig. 9 and Fig. 10.

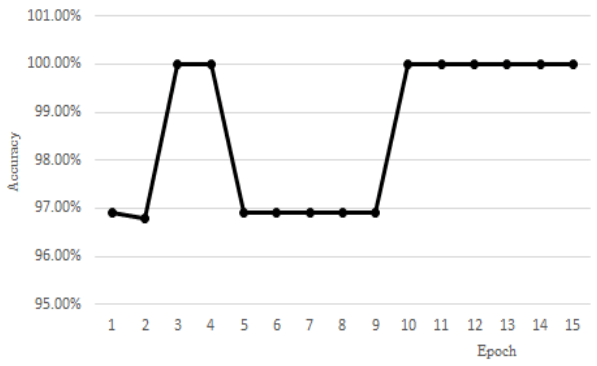


Figure 8. Curve of the accuracy change of the training set

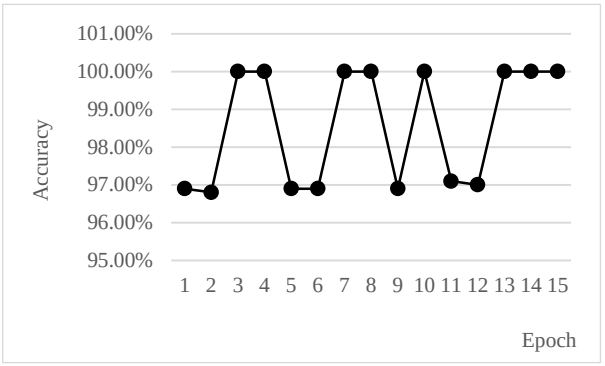


Figure 9. Curve of the change in the accuracy of the test set

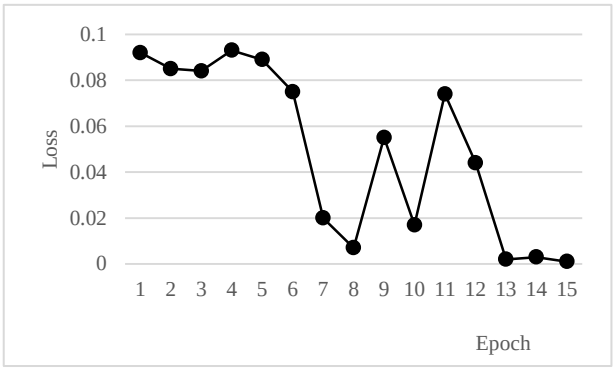


Figure 10. Loss change curve

This training will take the average accuracy of 10 independent experiments as the final experimental results, and the following table lists the recognition rates of six gestures.

TABLE I. GESTURE RECOGNITION RATE TRAINING EFFECT

Gesture category	Recognition rate
Digital gesture 1	95%
Digital gesture 2	99%
Digital gesture 3	96%
Digital gesture 4	83%
Digital gesture 5	86%
Digital gesture good	100%

From Table 2.1, it can be seen that the average recognition performance of the gesture recognition neural network model in this experiment reached 93.17%, which effectively recognized static gestures.

IV. CONCLUSION

Aiming at the problems of large number of parameters of gesture recognition model based on deep learning, slow training speed and high equipment requirements, which increases the cost, this paper proposes a gesture recognition system based on convolutional neural network. Using the template matching algorithm based on feature points, the gesture is finally effectively recognized by image acquisition, image preprocessing, gesture segmentation, etc. Experimental results show that the method can complete the recognition of static gestures, the recognition accuracy rate has been improved, and the recognition speed has been significantly improved, which has high real-time performance compared with other algorithms and has good generalization.

REFERENCES

[1] WENG H L, ZHAN Y W. Vision-based hand gesture recogni-tion with multiple cues [J]. Computer engineering & science, 2012, 34(2):123-127.

[2] L ü N, Yang Y J, Xu T. Sparse decomposition for data glove gesture recognition [C]. Proceedings of the 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). Piscataway, NJ: IEEE, 2017: 1-5.

[3] Pisharady P K,Vadakkepat P, Loh A P. Attention based detection and recognition of hand postures against complex backgrounds [J]. International Journal of Computer Vision, 2013, 101(3): 403-419

[4] Dai Y K, Zhou Z H, Chen X, et al. A novel method for simultaneous gesture segmentation an recognition based on HMM [C]. Proceedings of the 2017 International Symposium on Intelligent Signal Processing and Communication Systems. Piscataway, NJ: IEEE,November 6-9,2017: 684-688.

[5] SERMANET P, KAVUKCUOGLU K, CHINTALA S,et al. Pe-destrian detection with unsupervised multi-stage feature lear -ning [C]// 2013 IEEE Conference on Computer Vision and Pat-tern Recognition. Portland:IEEE, 2013:3626-3633.

[6] ZHANG C, ZHANG Z. Improving multiview face detection with multi-task deep convolutional neural networks [C]// 2014 IEEE Winter Conference on Application of Computer Vision. Steamboat:IEEE, 2014:1036-1041.

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.

[8] Ren S,He K,Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [C]. International Conference on Neural Information Processing Systems. [S.l.]: MIT Press, 2015: 91-99.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: unified,real-time object detection [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788.

[10] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector [C]. European Conference on Computer Vision. Springer International Publishing, 2016: 21-37.

[11] WU Yaoling. YCrCb color space face detection algorithm based on the design and implementation [D]. Chengdu:University of Electronic Science and Technology of China,2013.

[12] PENG Yaqin, CHENG Xiaogang. An optimized deep learning algorithm of convolutional neural networks [J]. Modern electronics technique, 2016, 39(23):179-181.

[13] SAXE A M, PANG W, KOH Z, et al. On random weights and unsupervised feature learning [C]// Proceeding of 2011 In-ternational Conference on Machine Learning. Bellevue:ACM, 2011:1089-1096.

Fine-grained Recognition of Ships Under Complex Sea Conditions

Jiaojiao Ma
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: m2578516632@163.com

Hong Jiang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 249479898@qq.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: yujun@163.com

Wei Li
Technological Center
Xi'an Technological University
Xi'an, 710021, China
E-mail: liwei@163.com

Haoqi Yang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: yanghaoqi@163.com

***Abstract*—For the traditional deep learning cannot solve the fog, coastal background interference, and the difficulty of small ships recognition, a multi-scale deep learning training model is proposed in this paper. Based on Faster R-CNN, this paper uses guided filtering to remove fog, as well as combined with negative sample enhancement learning to train the model, thus solving recognition of ship in complex sea conditions. And with multi-scale training strategy, the multi-scale ship samples are produced and sent to the network for training, so as to solve the problem of small target recognition. The experimental results show that compared with the Faster R-CNN, the precision and recall of our method increase by 6.43% and by 4.68% respectively. It solves the difficulty of ships recognition under complex sea conditions and small ship recognition that cannot be solved by traditional deep learning methods, the trained model has good generalization ability and robustness.**

Keywords- Ship Recognition; Complex Sea Conditions; Multi-scale; Fine-grained; Deep Learning

I. INTRODUCTION

Maritime target recognition is a key content in the field of computer vision. It has important practical significance for the automatic detection and recognition of ships [1-5], both in the civilian and military fields. With its broad application prospects, ship recognition has attracted significant attention all over the world. However, due to the problem of small size and changeable direction of ships, automatic recognition of ship in aerial images still faces many challenges. In addition, the complex sea condition also adds difficulty to ship recognition.

In recent years, with the rapid development of deep learning [6], the target recognition methods [7-11] based on deep learning has been widely used. Chen [12] et al used deep learning for the first time in the detection and recognition of ship targets. The image is searched for areas where ships may exist, and then analyzes candidate areas

to confirm whether or not they contain targets. The recognition rate is 91%.Bousetouane [13] uses the convolutional neural network to extract the features of the ship, and matches the ship with template in the template library. Then, the classification and recognition of the ship is performed, the precision is 89%.Shi [14] used the RBM which is important branch of deep learning to recognize ships, the precision is 95%.Wang [15] mainly analyzes the application of three kinds of mainstream methods (RCNN, Fast RCNN, Yolo) of deep learning about ship recognition, and compares the three methods to obtain the applicable scope of different methods. However, regardless of the above three methods or other existing methods, all of them cannot solve the following three problems: (1) it is impossible to solve the fog obstruction in practice; (2) the small ship cannot be recognized; (3) the target is smaller than the background area that results in the difficultly of training convergence.

II. RELATED WORK

This article carries on the detection and recognition of ship on the basis of Faster R-CNN, the purpose is to achieve the precise position and recognition of ship in the aerial image. The overall algorithm flow chart for ship recognition is shown in Figure 1.

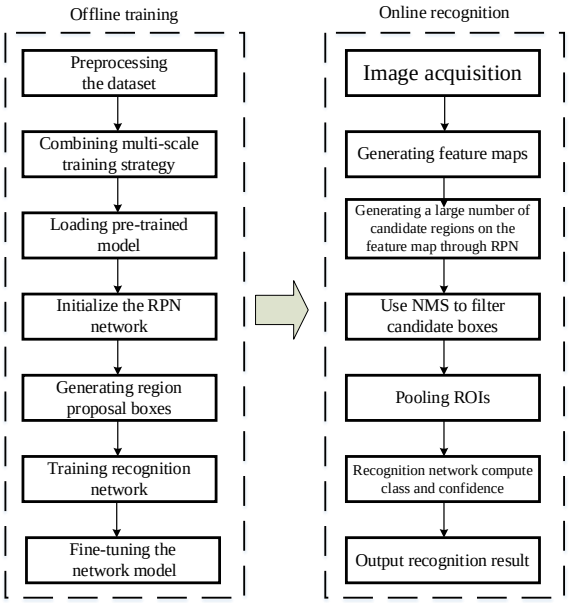


Figure 1. Algorithm flow chart for the ship recognition.

The whole process can be divided into two parts: offline training and online recognition .In the offline training process, there are mainly four steps: the first step is to preprocess the dataset ,to remove the fog in the image ;the second step is to join the multi-scale training strategy, the collected images are scaled to three scales training; the third step is to send the feature maps to the region proposal network(RPN) to generate the ship candidate regions; the fourth step, a multi task classifier is applied to target location regression and classification for ship candidate regions. During the online recognition, or testing process, we extracted the trained network model, tested it with real-time images, and output the location and class of ship to be recognized in the image, and analyzed the performance of the network.

A. Removal of Fog

As aerial image is often affected by complex weather, such as fog, this factor will cause blind to the ships and affect the recognition in subsequent experiments. The atmospheric scattering model [16] was proposed by McCartney in 1976 based on the atmosphere. The scattering model is an important theoretical basis for the removal of fog. The model is shown in equation (1).

$$I(x) = J(x)t(x) + A\left(1 - t(x)\right) \qquad (1)$$

In equation (1), $I(x)$ is an image with fog, $J(x)$ is an image without fog, $t(x)$ is atmospheric scattering rate, which reflects the ability of ray to penetrate the fog. A denotes atmospheric light intensity, and it is a constant vector.

In this paper, the method of guided filtering [17] is used to defog. The guided filtering obtains the result image based on the equation (1), including a guide image I, an input image P, and an output image q. The guide image I needs to be set in advance according to specific application, it can also be directly taken as the input image P. For the i^{th} pixel in the input image, the calculation method can be expressed as:

$$q_i = a_k p_i + b_k, \quad \forall i \in \omega_k \tag{2}$$

In equation (2), i is i^{th} pixel label, a_k and b_k are linear coefficients, and ω_k is local neighborhood. By calculating the gradients on both sides of equation (2), $\nabla q = a \nabla p$ can be obtained. When the input image p has gradients, the output q also has similar gradients, so that the feature of the ship edge can be preserved while defogging. The minimum cost function between the input image with fog and the output image is shown in equation (3).

$$E(a_k, b_k) = \sum_{i \in \omega_k} \left(\left(a_k p_i + b_k - p_i \right)^2 + \varepsilon a_k^2 \right) \tag{3}$$

In the above formula, we introduce ε constants to make a_k converge. We can know from equation (3):

$$a_k = \frac{\frac{1}{|\omega|} \sum p_i^2 - \mu_k \bar{p}_k}{\sigma_k^2 + \varepsilon}, b_k = \bar{p}_k - a_k \mu_k$$

Where μ_k and σ_k^2 are the mean and variance of the guide image p in the ω_k window, $|\omega|$ is the number of pixels in the $\bar{p}_k$ window, $\bar{p}_k$ is the mean of input image p in window ω_k.

For each pixel, the final relationship between the output image q and the input image p is shown as equation (4).

$$q_i = \frac{1}{|\omega|} \sum_{k, i \in \omega_i} \left(a_k p_i + b_k \right) = \bar{a}_i p_i + \bar{b}_i \tag{4}$$

In this paper, the fog image is used as the guide image to guide the filtering, so as to obtain the result image q after defogging, as shown in Figure 2.

(a) Original image (b) Result image

Figure 2. Result image after defogging.

B. Multi-scale Training

As aerial images are shoot at high altitude, the ship in images is relatively small so that affect the recognition accuracy. In addition, as samples are limited and the feature of ship cannot be fully extracted during training, which can result in false or missed detection. We use multi-scale samples, that is, set each image to multiple scales. The specific idea is to reduce the training samples to three scales (1024*1024, 512*512,256*256), as shown in Figure 3, it is equivalent to add large numbers of small ship samples for training in the training set. As the number of small target samples increase, that can ensure the network extracts the features of small target effectively, thus avoiding false or missed detections. The experiment shows that the multi-scale training can make the distribution of the target sizes of various classes of training more uniform, so that the trained network model can be more robust to multi-scale targets.

Figure 3. Multi-scale training sample images.

C. Feature Extraction

To achieve high recognition accuracy of ship, the effective features of the ship have to be extracted firstly, and the feature extraction of the ship is closely related to the structure of the convolutional neural network for feature extraction. The Convolutional Neural Network [18-20] (CNN) is a

feature extractor composed of convolutional layers and sampling layers. The advantage is that it does not require complex preprocessing. Feature extraction and pattern classification are completely put into a black box. Through continuous optimization to obtain the required parameters, CNN gives the desired classification at the output layer. The combination of convolutional layer and pooling layer will directly affect the effective feature extraction of ships. After a large number of experiments, the pre-trained ZF [21-22] network is chosen. Compared with AlexNet [23-24], GoogleNet [25-26], VGG [27-28], ResNet [29-30] and other networks, ZF has a low degree of structural complexity, and can also reduce the time complexity of feature extraction. The network structure of feature extraction is shown in Figure 4.

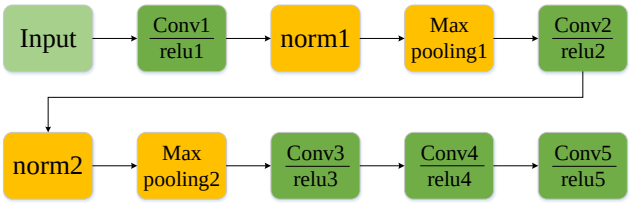


Figure 4. Feature extraction network structure.

D. Region Proposal Network

The accuracy of the ship recognition is directly related to the quality of the candidate region which is taken in the target detection algorithm. The algorithm combines the candidate regions with CNN. If we can find a way to extract only a few hundred of high quality ship candidate regions and have a high recall, we can not only promote the effect of target recognition, but also upgrade the processing speed of ship recognition. Region Proposal Network (RPN) can be a good solution to the above problem. RPN uses information such as textures, edges, and colors in the image to find out where the ship may exist in image. It can ensure that a higher recall can be maintained when fewer regions (thousands or even hundreds) are selected. This greatly reduces the time complexity of subsequent operations, and the candidate regions obtained are of higher quality than the target of the traditional sliding window method.

The structure of regional proposal network is shown in Figure 5.

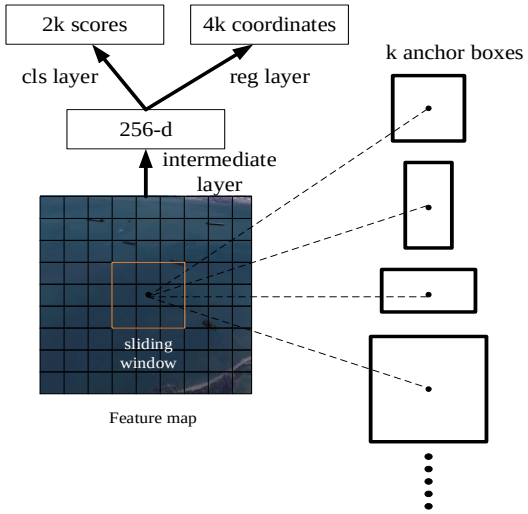


Figure 5. Region proposal network structure.

A fixed-size window slide is used above the feature map extracted by the feature extraction network. The center point of each window corresponds to k anchor points, and each anchor point corresponds to multiple sizes and aspect ratios of sliding windows (see Fig. 5). On the right side, RPN uses 3 sizes and aspect ratios, so each slide window has k = 9 anchor points. Correspondingly, in the position of each sliding window, nine regions are simultaneously proposed, so the classification layer outputs 2×9 parameters that reflects the probability of the region where ship is. The bounding box regression layer has 4×9 parameters, representing the vertex coordinate of the nine proposal regions.

E. Model Training

After using RPN, Faster R-CNN uses Fast RCNN to recognize and classify. Fast RCNN was proposed by Ross in 2015. This method solves the problems of RCNN detection slowly and training time consuming highly, and achieves end-to-end joint training. The RPN and Fast RCNN share the convolutional features in Faster R-CNN. Fast RCNN uses the high quality proposal regions provided by RPN, which greatly increases the speed of ship recognition. The schematic diagram of the Faster R-CNN is shown in Figure 6.

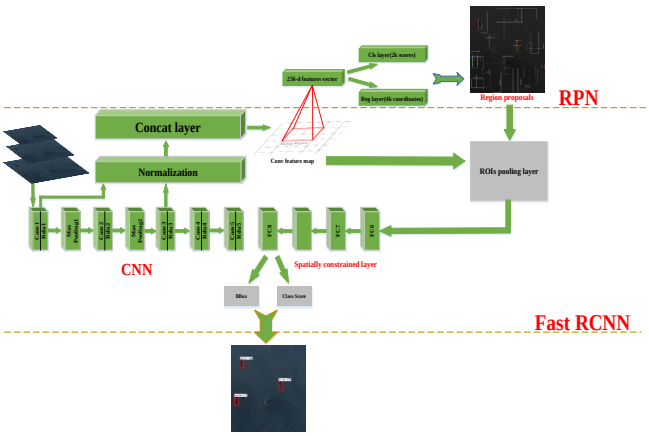


Figure 6. The architecture of proposed multi-scale Faster R-CNN for ship recognition. The simplified CNN model is surrounded by green boxes.

This article uses the RPN and Fast RCNN two-part network for ship recognition. Therefore, the back propagation algorithm cannot be used directly during the training process. Therefore, alternate training methods are used during training.

In the first stage, train RPN. The ImageNet pre-trained model (M0) is used to initialize the RPN network, and then the dataset is used to train the RPN network. After the training is completed, the model M1 is obtained.

In the second stage, train Fast RCNN. The ImageNet pre-trained model (M0) is also used to initialize the Fast RCNN network. Then the trained RPN network in first stage is used to obtain the proposal area P1, and the P1 is used to train the Fast RCNN network. After the training is completed, the model M2 is obtained.

In the third stage, train the RPN network again. Using M2 to initialize the RPN network and get the M3 network, this stage only fine-tunes the parameters of the RPN and sets the network parameters of ZF.

In the fourth stage, Fast RCNN is trained again .Use the M3 network to initialize the Fast RCNN, then use the third stage of the trained RPN network to obtain the proposal regions P2, and use P2 to train the Fast RCNN. This stage only fine tunes the parameters of the full connecting layer of Fast RCNN. In this way, both networks share the convolutional layer and form a unified network. The number of iterations for each stage is shown in Table 1.

TABLE I. FASTER R-CNN TRAINING PROCESS

Training stage	Network	Number of iterations
1	RPN	40000
2	Fast RCNN	40000
3	RPN	80000
4	Fast RCNN	40000

The algorithm adopted in this paper uses an end-to-end network for ship recognition, which avoids the trouble of buffering transfer data in the previous multi-stage training. It has greatly improved both in speed and accuracy, and achieved robust, rapid and accurate of ship recognition.

III. EXPERIMENT RESULTS AND ANALYSIS

A. Dataset

The dataset is provided by DataFountain and is composed of large numbers of ships photoed in bad weather, including four classes of ship: cargo ship, cruise ship, fishing ship, yacht. In order to show the recognition results better in the experiments, we call these four kinds of ships as huochuan, youlun, yuchuan and youting. Parts of the sample image are shown in Figure 7:

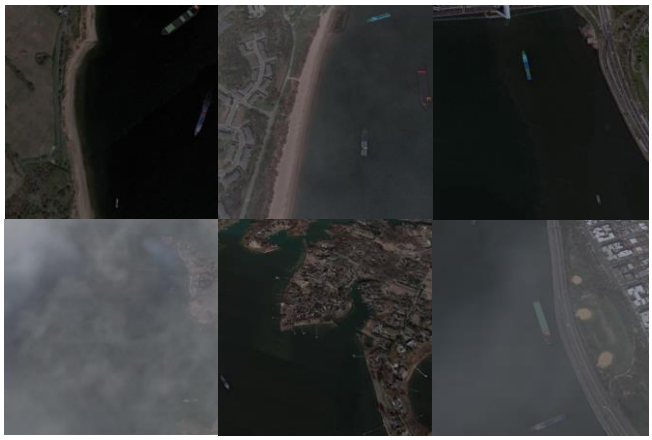


Figure 7. Part of the sample images (huochuan is cargo ship, youlun is cruise ship, yuchuan is fishing ship, youting is yacht).

In the experiment, the dataset consisted of 33,397 images, of which the number of huochuan is 33756, the number of youlun is 8028, the number of youting is 13,608, and the number of yuchuan is 10,120.In addition, 5000 images with fog and coastal background were selected as nega-

tive samples of another two classes .As in deep learning, the background is also a class in the target recognition, so this experiment is set to 7 classes.

Since the recognition model adopted in this experiment is Faster R-CNN, the location of the ship must be marked in the training set firstly. That is, finding the region of interest (ROI), and setting the ground-truth of ships in the image, and mark the specific class of ship. The ROIs of some training samples are shown in Figure 8.

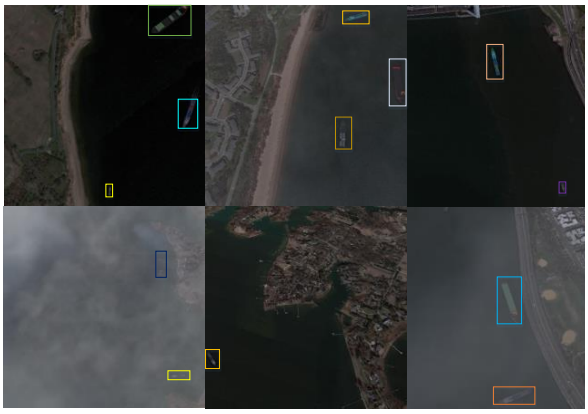


Figure 8. The ROIs of some training samples.

In order to evaluate the effectiveness of the algorithm, this paper uses precision and recall to measure the performance of the model. Both the recall and the precision range are between [0,1]. The calculation equations are as shown in (5) and (6):

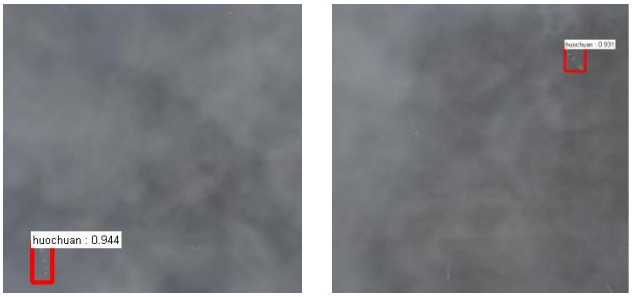
$$precision = \frac{TP}{TP + FP} \tag{5}$$

$$recall = \frac{TP}{TP + FN} \tag{6}$$

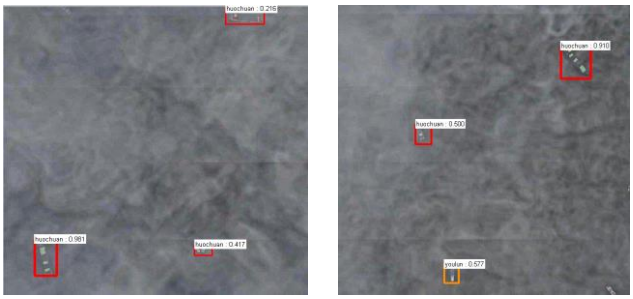
In the above formula, TP is the number of samples which is recognized correctly, FP is the number of false recognition, and FN is the number of missed recognition.

B. Comparison of Recognition Result After Defogging

In order to verify the removal of fog algorithm used in this paper is practical and effective, some samples with fog are selected from the dataset for testing. The experiment results are shown in Figure 9.



(a) Recognition results without defogging



(b) Recognition results after defogging

Figure 9. Comparison of ship recognition experiment with fog.

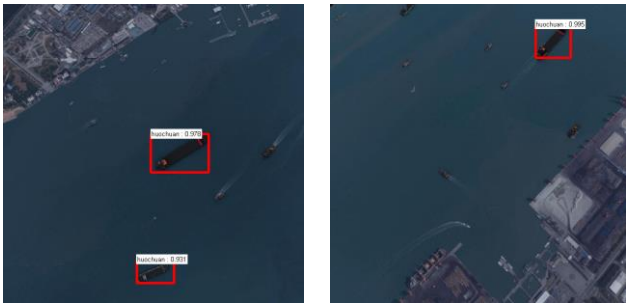
From the above experiment results, we can see that the method used in this paper not only can defog effectively, but also can retain the effective features of the ship as much as possible, and eliminate the unfavorable factors for ship recognition.

C. Comparison of Two Algorithms in the Same Sea Condition

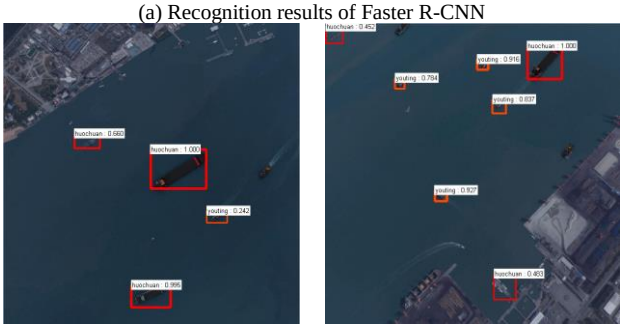
To further verify the efficiency of the proposed algorithm, the algorithm is compared with Faster R-CNN, 100 images (including 341 ships) are selected from the training set for testing. The precision and recall of the two algorithms are shown in the table 2.

TABLE II. COMPARISON OF RECOGNITION EFFICIENCY OF THE TWO ALGORITHMS

Detection Method	TP	FP	TN	precision/%	recall/%
Faster R-CNN	297	40	44	88.13	87.1
Our	313	18	28	94.56	91.78



(a) Recognition results of Faster R-CNN



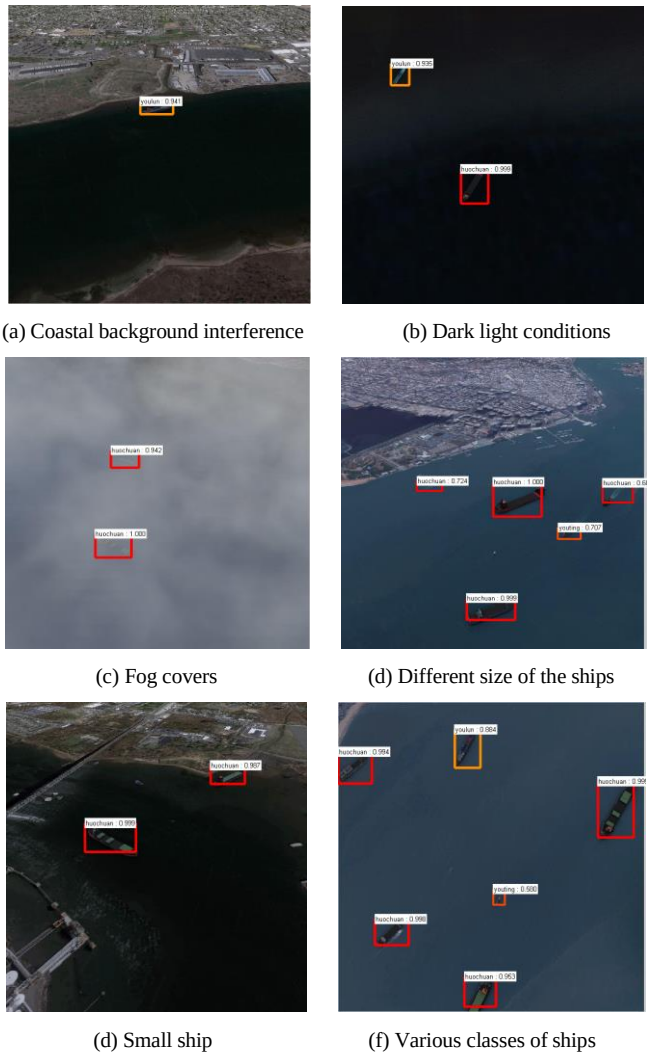
(b) Recognition results of our method

Figure 10. Comparison of two algorithms in the same sea state.

From Table 2 above, it can be seen that the model trained by our method is higher than the model trained by Faster R-CNN in terms of precision or recall. Compared with Faster R-CNN, the precision is increased by 6.43%, the recall is increased by 4.68%. It can recognize the various classes of ship in the images. From Figure 10, it can be seen that both the precision and confidence of ship recognition have increased significantly.

D. Recognition Results Under Various Sea Conditions

In order to verify the robustness of the proposed algorithm, large numbers of images that contain various ships under different sea states were selected from the dataset for testing. Its recognition result is shown in Figure 11.



(a) Coastal background interference (b) Dark light conditions

(c) Fog covers (d) Different size of the ships

(d) Small ship (f) Various classes of ships

Figure 11. Recognition results under various sea states.

It can be seen from Figure 11, our method can deal with the interference of the coastal background as shown in (a), (d), (e), the fog cover as shown in (c), the size of the ship differs greatly as shown in (d), the small ships as shown in (e), and the various classes of ships are as shown in (f), all have better recognition results. The experiment results show that the proposed algorithm is widely applicable to the ship recognition under complex sea states, and the accuracy and real-time performance meet the actual requirements, and it has strong robustness.

IV. CONCLUSIONS

The purpose of this paper is to solve the difficulties of ship recognition and small target recog-

nition under complex sea conditions. Firstly, through the guidance filtering to remove fog and combined it with the negative sample reinforcement learning method to train model, the problem of the influence of the complex sea situation is solved. Then, combined with multi-scale training strategy, different sizes of ship targets are sent to the network for training, thus solving the problem of small target ship recognition. Experiments show that this algorithm can extract the ship features of different sizes and conditions. The network model has strong robustness to dark light condition, fog cover, different size of ships and interference of coastal background, and the processing time of each image is at millisecond level that can meet the requirement of real time.

REFERENCES

[1] W. Wang, Y. Ji, X. Lin, "A novel fusion-based ship detection method from Pol-SAR images," Sensors, vol. 15, pp. 25072-25089, Sep 2015.

[2] H.-L. Wang, M. Zhu, C.-B. Lin and D.-B. Chen, "Ship detection in optical remote sensing image based on visual saliency and AdaBoost classifier," Optoelectronics Letters, vol. 13, no. 2, pp. 151–155, 2017.

[3] M. Kang, X. Leng, Z. Lin, K, Ji, "A modified faster R-CNN based on CFAR algorithm for SAR ship detection," Proc. Int. Workshop Remote Sens. Intell. Process. , pp. 1-4, May 2017.

[4] Yao. Y, Jiang. Z, Zhang. H, Zhao. D, Cai. B, "Ship detection in optical remote sensing images based on deep convolutional neural networks," J. Appl. Remote Sense, Nov 2017.

[5] C. Schwegmann, W. Kleynhans, B. Salmon, "Synthetic aperture radar ship detection using Haar-Like features," IEEE Geosci. Remote Sens. Lett., vol. 14, no. 2, pp. 154-158, Feb. 2017.

[6] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, vol. 521, pp. 436-444, May 2015.

[7] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Proc. IEEE Conf. Comput. Vis. Pattern Recog, pp. 580-587, 2014.

[8] R. Girshick, "Fast R-CNN," Proc. IEEE Conf. Comput. Vis. (ICCV), pp. 1440- 1448, Jun. 2015.

[9] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi,"You only look once: Unified real-time object detection," Proc. IEEE Conf. Comput. Vis. Pattern Recogn., pp. 779-788, 2016.

[11] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, "SSD: Single shot multibox detector," Proc. Eur. Conf. Comput. Vis., pp. 21-37, 2016.

[12] Chen Liang, Wang Zhiru, Han Zhong, et al, "Ship Target Detection and recognition method based on visible remote sensing image," Science & Technology Review,pp.77-85, 2017.

[13] F.Bousetouane, B.Morris, "Off-the-shelf cnn features for fine-grained classification of vessels in a maritime environment," Advances in Visual Computing, pp. 379-388, 2015.

[14] Shi Danrong, "Ship detection of SAR image based on deep learning," Xidian University, 2015.

[15] Wang Bing, "Research on ship detection Based on deep learning," Xiamen University,2017.

[16] E. J. McCartney, "Optics of the Atmosphere: Scattering by molecules and particles," John Wiley and Sons, 1975.

[17] K. He, J. Sun, X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 6, pp. 1397-1409, Jun. 2013.

[18] N. Ketkar, "Convolutional neural networks in Deep Learning with Python,"Springer,pp. 61-76, 2017.

[19] P. Kim, "Convolutional neural network in MATLAB Deep Learning," Springer, pp.121-147, 2017.

[20] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, "Fully convolutional neural networks for remote sensing image classification," Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS), pp. 5071-5074, Jul. 2016.

[21] M.A. Rafique, W. Pedrycz, M. Jeon, "Vehicle license plate detection using region-based convolutional neural networks, "Soft Comput., pp. 1-2, 2017.

[22] M. D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks," European conference on computer vision, pp. 818-833, 2014.

[23] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Proc. 25th Int. Conf. Neural Inf. Process. Syst., pp. 1097-1105, 2013.

[24] J. Hosang, M. Omran, R. Benenson, B. Schiele, "Taking a deeper look at pedestrians," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4073-4082, 2015.

[25] C. Szegedy et al., "Going deeper with convolutions," Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1-9, 2015.

[26] Z. Zhong, L. Jin, Z. Xie, "High performance offline handwritten Chinese character recognition using GoogLeNet and directional feature maps," Proc. 13th Int. Conf. Document Anal. Recognit., pp. 846-850, 2015.

[27] A. Vedaldi, K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," Proc. 23rd Annu. ACM Conf. Multimedia Conf., pp. 689-692, 2015.

[28] T. Sercu, C. Puhrsch, B. Kingsbury, Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR," Proc. IEEE Int. Conf. Acoust. Speech Signal Process, pp. 4955-4959, Mar. 2016.

[29] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, pp. 770-778, 2016.

[30] J. B. Zhang, Y. Zheng, D. K. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," Proc. 31st AAAI Conf. Artif. Intell., pp. 1655-1661, 2016.

A Survey of Ternary Optical Computer Research

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

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

***Abstract*—TOC, as a new type of computer, uses the state of light: it uses the non light state and two polarized light states with orthogonal polarization directions to represent information, and uses a rotator and polarizer to change the three light states, thus completing ternary logic operations. This new type of computer has the advantages of numerous processor bits, independent use of processor bits in groups, and reconfigurable computing functions of processor bits; After more than 20 years of development, we have achieved fruitful results. In this paper, the research progress of ternary optical computer is comprehensively and thoroughly analyzed and summarized, from the concept of TOC theory to the kilobit processor being studied so far. Firstly, the theory of ternary optical computer is summarized. Secondly, according to the core theory of ternary optical computer and the achievements that have been made, ternary optical computer has also made some progress in software and data calculation, such as operation result file generation software, addition routine, multiplication routine, etc. Finally, the future research directions of ternary optical computer are analyzed and prospected. Firstly, the structure of ternary optical computer is introduced, and the existing research results are summarized. Finally, the future research directions of ternary optical computer are analyzed and prospected.**

Keywords-Ternary Optical Computer; Ternary Logic; Operation-data File

I. INTRODUCTION TO TERNARY OPTICAL COMPUTER

A. Background introduction

With the development of electronic computers and the advent of the 5G network era, people cannot live without processors in their daily lives, such as general-purpose processors, digital signal processors and graphics signal processors on mobile phones; Computer systems can solve some complex engineering problems. Due to the widespread use of processors, the computing requirements are also getting higher and higher. In order to break through the limitations of too low information transmission speed and too few data bits in electronic computers, scientists in the computer field are gradually trying to develop a faster, more reliable The ternary optical computer is one of the new computers with higher, required equipment and less power consumption. The ternary optical computer has a processor number of up to one million. Part of it can serve an application program independently, the hardware computing function of each processor bit can be reconfigured at any time, and its parallel adder can be completed in 3 instruction cycles when calculating data of any number of digits, maintaining traditional programming technology and with the advantages of very little power consumption, it

DOI: 10.2478/ijanmc-2022-0036

has become the best choice for connecting electronic processors today.

The ternary optical computer uses the two mutually perpendicular polarization directions and the non-optical state of the optical state to represent information, uses the two orthogonal polarization directions of the optical state and the non-optical state, uses the liquid crystal to change the polarization direction of the light, and uses the polarizer. Check the polarization direction, use liquid crystal and polarizer to realize the transition from light state to no light state, and then realize many three-valued logical transformations. The physical characteristics of light determine the "ternary value" in the ternary optical computer; the same as the binary electronic computer, it follows the same physical principle, but the ternary optical computer has higher coding efficiency.

B. Conceptual diagram of ternary light computer

As the core device of the ternary optical computer, the ternary optical processor adopts a special construction strategy - based on the electronic computer, "optical calculation, electrical control" to construct the ternary optical computer.

The conceptual diagram of the ternary optical computer is shown in Figure 1: it consists of three main parts: the control system, the data input/output system and the optical processor. Among them, the part marked by the circle is the ternary optical processor; the thick line between the input/output array and the optical processor represents the thousand-bit data line, and the thin line between the m-channel data input server and each input storage array represents the 64-bit data Wire, The thin lines of the m-channel data output server and each output storage array represent the 64-bit result output line, The line sent out by the control system is the control line, including the reconfiguration command control line for the arithmetic unit and the control line for each electronic component.

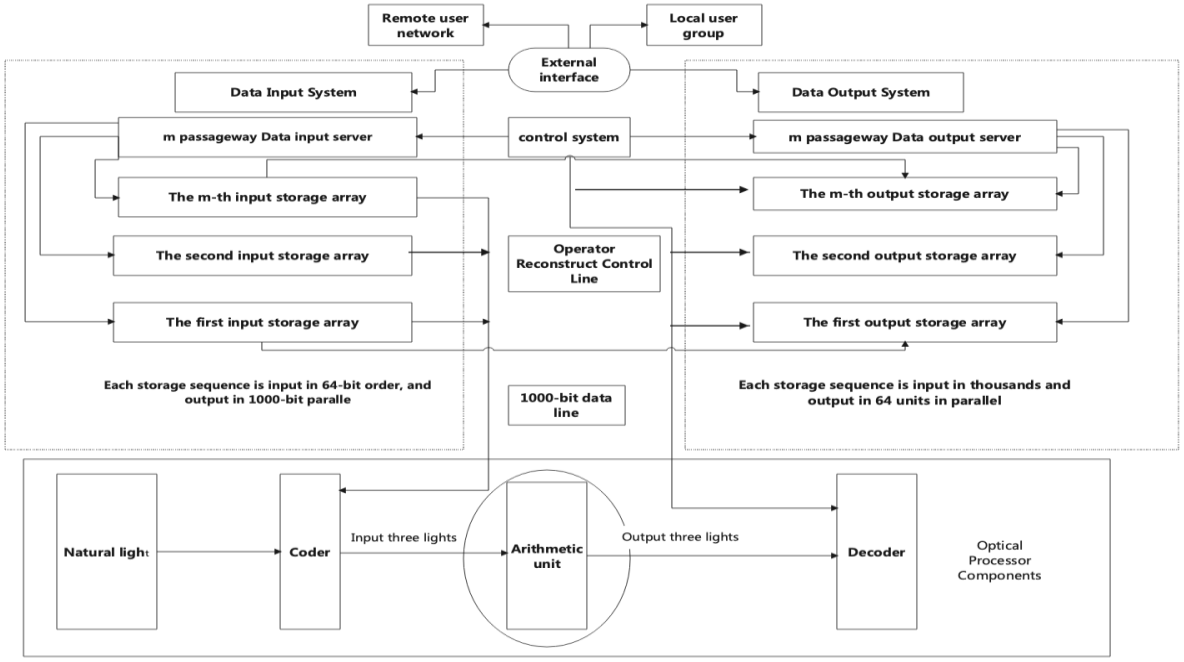


Figure 1. Conceptual diagram of ternary light computer

At present, this conceptual diagram is used as the basis for all ternary optical computers, and its working principle is as follows: the user of the ternary optical computer submits the computing task to the ternary optical computer system through the external interface, the control system receives the sent commands, and the data input The system receives the sent data, and then the control system generates control commands for each component according to the user's command: the controller applies for a data channel to the data input system for the user's input data, and the m-channel data input server allocates a certain storage array. After the array is full, The data is sent to the encoder in the optical processor unit through the kilobit data line, and then the input data storage array informs the control system; The controller sends out the operator reconstruction control command to the operator, the operator processes the data efficiently with ternary operation, and then sends the operation result to the decoder; Then output the result to an output storage array of the output system; finally, after the output storage array receives the calculation result, it sends the result to the m-channel data output server through a 64-bit result output line, and returns it to the user through the external interface.

II. ACHIEVED RESEARCH RESULTS

The ternary optical computer has been developed for more than 20 years. The development process is shown in Figure 2.

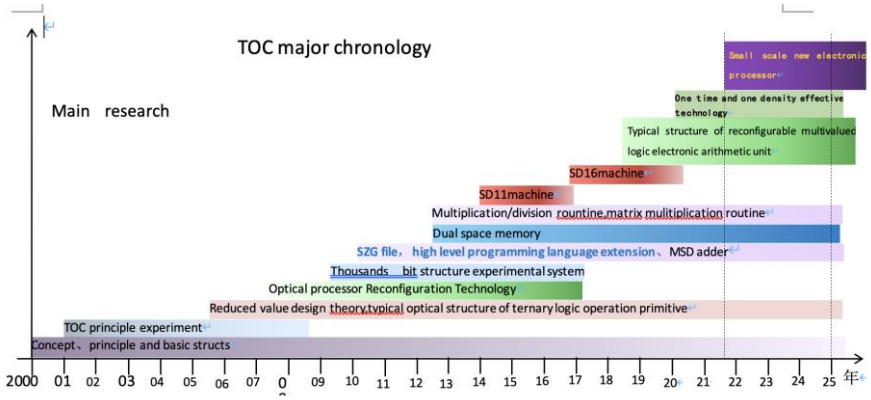


Figure 2. Development history of ternary optical computer

A. *The basic principle and origin of ternary optical computer*

In 1999, Professor Jin Yi discovered that two polarized states with orthogonal polarization directions can be used, the light state and the non-light state, the polarization direction of light can be changed with liquid crystal, the polarization direction can be checked with polarizer, and the light state can be changed from light state to non-light state with liquid crystal and polarizer. The transformation of the optical state, and then realize the three-valued logical transformation.

At the beginning of 2000, Professor Jin Yi carried out a series of principle experiments, which proved the feasibility of building a ternary optical computer, and also realized that the ternary optical computer would have many digits for calculation. But in this period, there is no clear method for how to convert the non-light state into the light state, I just wanted to use the interchangeability of symbols in logical operations to avoid generating an operator with a light output from two non-light inputs. Later, an

auxiliary light source was used to complete this kind of operation. In today's ternary optical processor prototype system, this auxiliary light source is implemented with a liquid crystal pixel,

In 2002, Professor Jin formally proposed the principle of ternary optical computer and the hybrid structure of optoelectronics in his doctoral dissertation, established the prototype of the theoretical system of ternary optical computer, and constructed the overall structure of ternary optical computer and the optical structure of main components, designed. The realization light path of the three-valued logic operation generated by the general logic student. At the same time, the concept of calculating the position and calculating the road was established. Based on this, the idea of the ternary optical computer distance management was established, the original characteristics of the optical computer software system were pointed out, and the ternary optical computer was extended to optical fiber communication. In the field of optical fiber communication, the idea and implementation scheme of using ternary optical signal for optical fiber communication are proposed.

B. The Proposition of Carry Direct Adder Idea

In August 2004, Jin Yi, He Huacan and others created the basic principle of the carry-to-parallel ternary optical computer adder, and created the carry-to-carry adder.

When a ternary computer builds an adder, it always identifies the addition operation as a combination of binary logic operations, and then uses a binary logic operator to build the adder. In electronic computers, in order to solve the single-line delay when carrying Carry-first strategy provides a theoretical basis for the design and implementation of a full adder in a ternary optical computer.

C. Balanced ternary numerical representation

Traditional electronic computers use "packetized parallel carry chains"; this approach increases the speed of operations by increasing the complexity of the use case, and the complexity of the computer increases with the number of bits in the computer. For an optical computer with a huge number of bits, this design will increase the structure of the optical computer adder.

In 2006, the ternary optical computer team introduced the balanced ternary numerical representation and calculation method into the ternary optical computer for the first time, which provided a theoretical basis for the study and realization of the arithmetic components of the ternary optical computer.

In 2007, Yin Xunwei, a master's student, studied the technology of "elimination of photoelectric converters" and used this technology to effectively simplify the structure of the half-adder, reduce the engineering difficulty of the half-adder, and shorten the working cycle of the half-adder.

In the ternary optical computer, the symmetric ternary {-1, 0, 1} is used to represent the data. Under the symmetric ternary representation, the positive and negative numbers have the same expression form, so the addition and subtraction operations have the same The operation process of, which also brings convenience to the fact that the subtraction operation does not need to be complemented; the negative number can be directly sent to the adder to complete the subtraction operation. Although the adder in the symmetric ternary representation still has the carry delay problem, it simplifies the subtraction operation.

D. Encoder/Decoder Experimental Research

In 2007, Yan Junyong and Sun Hao of the ternary optical computer team conducted an experimental study on the one- and nine-bit encoders and decoders of the ternary optical computer using separate optical components. This experiment proved that the ternary optical signal encoder and the feasibility of the decoder, the ternary optical computer has since entered the experimental stage,

E. Research on Data Storage of Ternary Computer

In 2007, PhD student Zuo Kaizhong discussed the problem of polarized light data storage and proposed a ternary polarization holographic digital optical storage method based on bacteriorhodopsin film. In the same year, a ternary polarization holographic digital storage method based on in dole fulgine anhydride film was proposed. This method can realize parallel addressing and read and write operations in units of pages, which is the beginning of the development of ternary optical computer optical memory.

F. Carry out research on data bit management of ternary optical computer

In 2007, aiming at the high data width (number of data bits) of the ternary optical computer, on the basis of the proposed four data bit management strategies and concepts such as bit count, count track, count bit type, count track type, etc. Professor Yi further enriched and embodied the basic strategy of high data width management, and made the initial planning for the design of the ternary optical computer operating system.

G. Propose the theory of devaluation design

In 2007, Ph.D. student Yan Junyong came up with the idea of finding the most basic unit in the structure of a ternary optical processor, and carried out the earliest research and achieved the initial results. Later, Professor Jin Yi, Yan Junyong and Zuo Kaizhong doctoral students focused on this work. After nearly a year of in-depth research, they discovered the law of depreciation design, and extracted the depreciation design theory from this law. Among them, Dr. Yan Junyong established a corresponding design specification for depreciation, and applied it to the components of the three-valued logic operator. By combining the theory of value design and multi-valued logic, he proposed a two-level design with liquid crystal and polarizer as the main components. A general implementation structure of a meta-ternary logic optical operator.

H. Carry out research on ternary optical computer application

In 2007, under the guidance of the reduced-value design theory, a ternary logic optical processor was implemented, and a ternary optical computer experimental system with the processor as the core device was completed. The three-valued logic optical processor adopts the two-dimensional plane information processing method, which has unique advantages: it has huge data bits, reconfigurable logic operation capability, and realizes three-valued operation characteristics. Therefore, it is completely possible to find suitable application points for the optical processing system, and in-depth exploration can give full play to the system.

Application research of various advantageous properties. This kind of research results will be directly transplanted to the new platform after the successful research of the thousand-bit reconfigurable ternary logic optical processor.

I. Experiment system

In 2011, Dr. Ouyang Shan designed the thousand-bit reconfigurable parallel TOC

experiment system SD11, which was the first practical TOC. Using the related technology of ternary optical computer, a prototype structure of the ternary optical computer monitoring system is established, which gives the initial idea of the ternary optical computer monitoring system.

SD11 experimental system provides the basis for future multiplication routines, monitoring systems and task management, and TOC is officially entering the experimental stage.

J. Ternary light computer monitoring system

In 2011, Dr. Wang Xianchao realized the first prototype of a ternary optical computer monitoring system with effective management of optical computer processor resources with huge data bits as the main content.

After half a year of research by Dr. Wang Xianchao, he finally determined the monitoring architecture of TOC, as shown in Figure 3. The workflow is as follows: The main function of the client is to input the truth table according to the user (about the input format of the truth table, the operand Carry out communication conversion with the truth table to generate the data represented by the internal code of communication, generate the operation request of the user, and send it to the server. On the other hand, when receiving the operation result represented by the internal code of communication, it is communicated The inverse transformation generates the operation result represented by ASCII code ; the main function of the server side is to first sort out the operation request received from the client, and insert the sorted operation request into the operation request linked list according to its priority; The operation request is scheduled according to a certain scheduling strategy, and the newly scheduled operation request is sent to the lower computer; after the scheduling is completed, the processor resources are allocated to each operation request, and the reconstruction code of the operator required for each operation is obtained at the same time, and the processing The processor allocation result and reconstruction code are sent to the lower computer; the lower computer completes the operation for the user according to the operands in the user operation request, the processor allocation result and the operator reconstruction code; finally, the operation result is sent to the client.

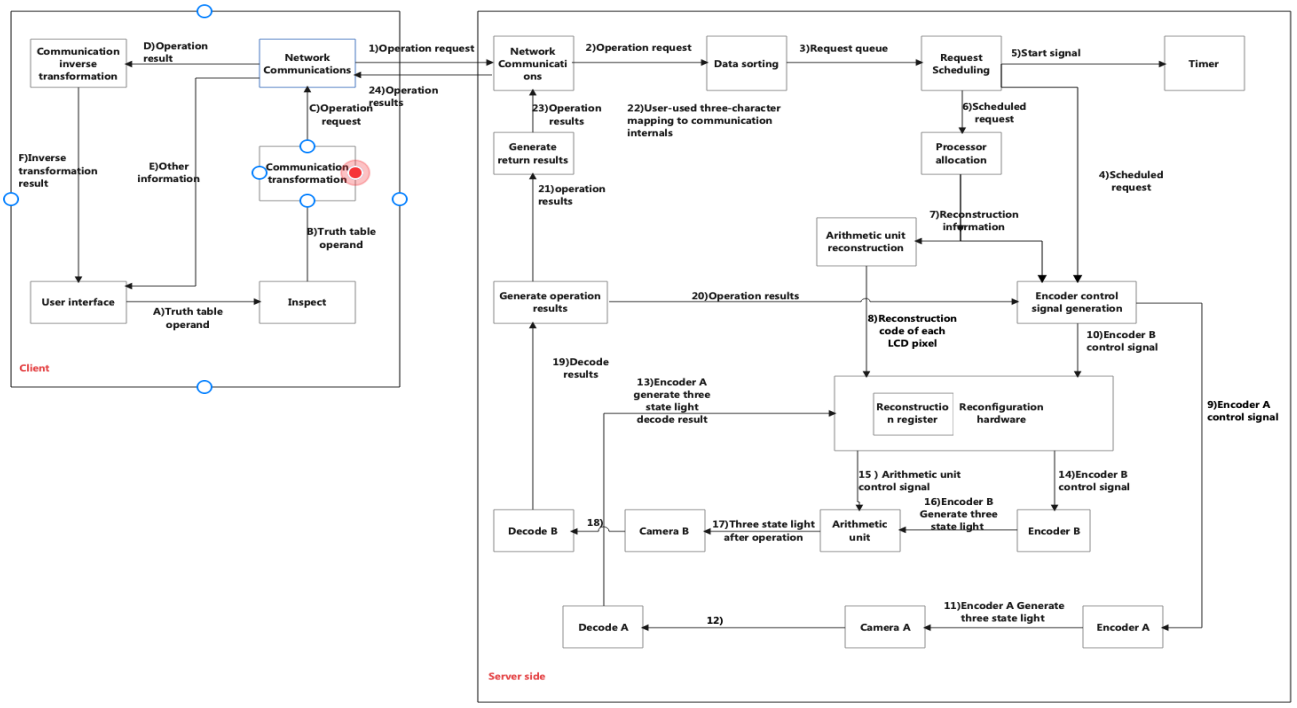


Figure 3. TOC Monitoring Architecture

Dr. Wang Xianchao conducted a series of tests on it, verified its robustness, reliability and correctness, proposed an algorithm for proportionally allocating optical processor resources of a ternary optical computer, and implemented the corresponding management program. The practical significance lies in enriching the theory and technology of data bit management in computer monitoring system. Its practical significance lies in paving the way for users to use ternary optical computer conveniently and efficiently. The use of ternary optical computers to solve challenging major problems can fully reflect the social and economic value of TOC, design and build the monitoring system software of ternary optical computers, directly use ternary optical processors for high-level programming languages , and for early Lay the foundation for the formation of a practical ternary optical computer monitoring system.

K. Operations - Data Files

In order to maintain the original programming habits and better utilize the advantages of TOC in the application, the team members use the form of operation-data files (SZG files) to include all TOC features. Therefore, the main way to use TOC is the SZG file.

In 2012, Song Kai established the first version of the SZG file, which mainly established that the SZG file was composed of two parts: the header for recording the overall information and the data area for recording the original data; the operation-data file for the ternary optical computer was established. The expression form is *SZG, and the corresponding calculation result is recorded as *_R.SZG, where the wildcard * is the full name of the original data file (including the storage path) given by the user. Based on this

version of the SZG file, the C language hungry MPI is completed. Basic expansion of the statement. From the experiment, it is clear that two kinds of computers can work together in the same application program by extending the existing programming language based on the SZG file to form the malicious application program platform of TOC.

2013, the team Jin Yi and others established the second version of the SZG file. This version of the SZG file defines the header form of the SZG file in more detail, and gives the names and functions of each item contained in the file header. and the number of bits, giving the number and location of reserved bits. This version not only makes SZG files have basic practical value, but also lays the foundation for the construction of TOC task management software, data bit allocation software, processor reconstruction software and other underlying core software.

With the development of TOC underlying software research, more detailed definitions are required for each item in the SZG file header to ensure that the TOC underlying core software can unambiguously understand the user's intention expressed in the SZG file. With the establishment of 40 -bit multiplication operation routines, division operation routines, FFT and DFT algorithm routines of TOC, SZG files are required to express more complex operation requests, and more mathematical operation routines are developed for third parties. Create conditions with common algorithm routines. With the deepening of the exploration of the potential application scenarios of TOC, it is also required that the original data or operation results can be easily extracted from the existing SZG files or their result files to generate new SZG files - forming a SZG file chain. Therefore, the third version of the SZG file format was established in 2015, which is also the finalized version, as shown in Figure 4.

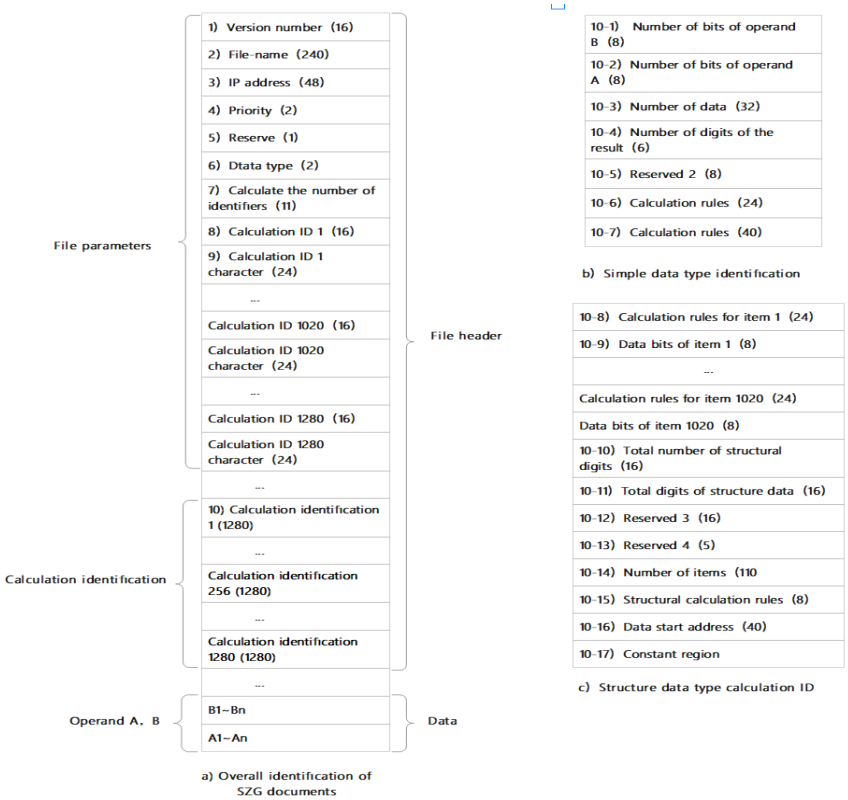


Figure 4. Operations - Data File Format

SZG file is the only way for TOC to obtain user operation data. SZG file consists of two parts: file header and operation data (file body). The file header records the structural parameters of each operator required by the user and the overall parameters of the operation data. The information in the file header mainly provides parameters for reconstructing the optical processor. The data area is the operand input by the user, including operand b and operation Number a. In order to ensure an accurate understanding of the user's operation data, the form of the SZG file constructed by the user must be consistent with the form of the SZG file read by the TOC. Therefore, the SZG file format definition is the core of this technology.

L. 40 - bit multiply routines

Multiplication is one of the important functions required in TOC's march towards the field of numerical computing. In 2013 , a postgraduate student Hu Xiaojun designed and implemented a 40- bit multiplication routine for ternary optical computer by taking advantage of the features of TOC, which is reconfigurable, with a large number of data bits and no carry delay. , through the M transformation in the three-valued logic to generate the partial product of the intermediate process of multiplication, and then use the iterative calculation method of adding two pairs to sum the values to obtain the final product. Hu Xiaojun established a specific operation routine by adopting the single M operator scheme and the "sum number" addition iteration scheme of the pipeline calculation combined with the design and planning of the TOC monitoring system, and gave the specific method for the upper-level users to use this routine; the first implementation of TOC The multiplication function is a key arithmetic operation function of TOC, and it also provides research ideas and method comparisons for the future development of other functions of TOC (such as matrix multiplication).

M. Ternary light computer physical machine

In August 2016, when optimizing the structure of the ternary optical processor and reconstructing the system, team member Dr. Li Shuang proposed the structure of the ternary optical computer - SD16 structure, and built SD16 - No.0 machine in November 2016 (Figure 5).

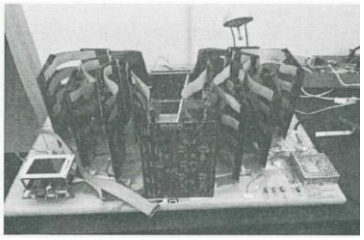


Figure 5. SD 16 - No. 0 machine

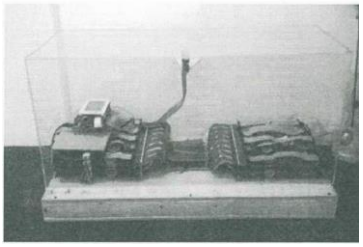


Figure 6. SD 16 - No. 1 machine

September 2017, in order to reproduce the structure of SD 16 and at the same time optimize the assembly method of each component, the team built SD16 - No. 1 machine (Figure 6).

SD 16 physical machine structure is a leap in the development of TOC, which proves the authenticity of TOC, and also means that all theories are based on realization, providing an experimental environment for all subsequent researches.

N. Computational data file generation software

Objectively speaking, after defining the final version of the SZG file format, when the user is familiar with the defined format, any editing software can be used to generate the SZG file from the original protective gear, but this method increases the calculation time, so in 2018, Dr. Li Shuang designed the SZG file generation software. This software takes up a small amount of system resources, does not increase the burden on the client system, avoids data loss, stores the user's data in non-volatile memory, and ensures that the existing SZG files are retrogradely tracked and modified at any time, which is convenient for users to update themselves.

The function of the software is to generate the initial SZG file through the user 's input of original data and calculation rules. Considering the friendliness of the user's input interface, the correctness, completeness and simplicity of the user's input information determine the final determination of this document as shown in Figure 7. As shown in the user input interface of the SZG generation software, the user only needs to input data through the prompt, and then the software will generate the file operation, which greatly improves the computing efficiency.

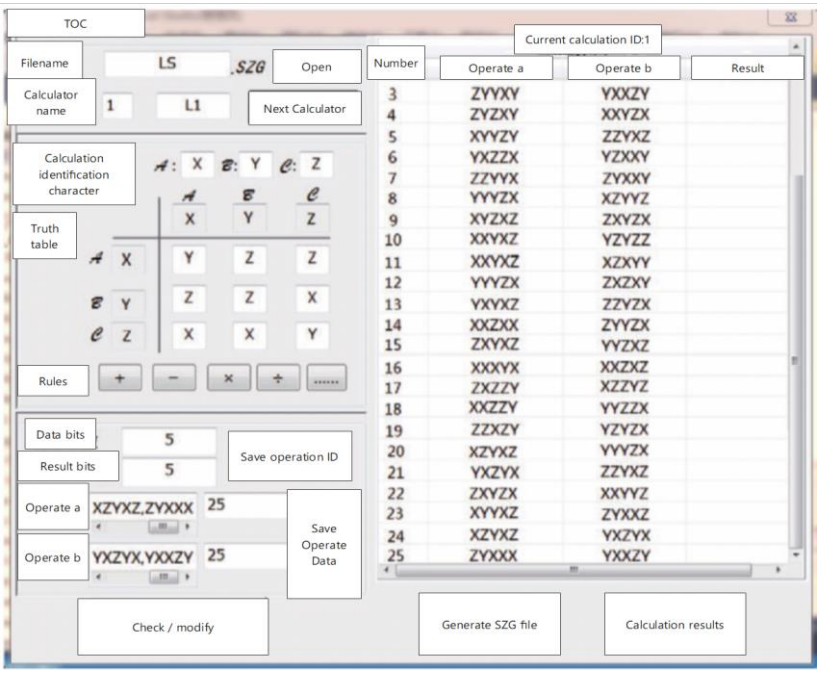


Figure 7. User Input Interface

The SZG file software generation software provides a corresponding programming platform for TOC, gives the basic theory and core technology of building a new computer programming environment, improves the existing theoretical foundation of TOC, and prepares for TOC to enter the application stage.

O. Big task management software

In2018, Dr. Zhang Sulan studied the following contents based on the characteristics of the ternary optical processor with a large number of processor bits, which can be grouped and allocated bit by bit:1) Before the ternary optical processor starts to calculate the specific data, it needs to be Reconstruct various calculators

required by the user; 2) The original data is output to the reconstructed calculators in a simple and fast way. To this end, the research team proposed a calculation-data programming model, and then structured the ternary optical computer software system, classified the calculation tasks proposed by users, and determined the control objects of the large task management software.

After nearly 5 years of hard work, the large task management software developed has the following functions: correctly accept the user's computing request; generate corresponding tasks according to the calculation strategy, efficiently manage and schedule these tasks; reasonably allocate the received large tasks Ternary optical processor bit resources; and generate reconstructed instruction frames and operand encoding frames that meet the requirements of the ternary optical processor format; finally, the operation results of the decoder should be obtained and correctly returned to the corresponding users.

The design and implementation of the large task management software provides ideas and experimental basis for the detailed classification of the subsequent TOC tasks. In fact, the implementation of the large task management software only completes a small part of the TOC task classification, and the remaining classification needs to continue. Research,

III. FUTURE RESEARCH DIRECTIONS

In 2007, under the guidance of the reduced value design theory, the ternary logic optical processor was realized, and the ternary optical computer experimental system with the processor as the core device was completed. At that time, the experimental system could not solve the practical scientific application problems. Until the advent of the first thousand-digit SD 11 experimental system in 2011, it paved the way for subsequent TOC time studies. Subsequently, the birth of the TOC prototype SD 16 in 2016 also verified the theory of the three-week optical computer again, and also carried out related research: the realization of cellular automata computing mode and the realization of optical vector-matrix multiplication.

In recent years, small-scale new electronic processors and reconfigurable multi-valued logic electronic operators are also being studied.

IV. CONCLUDING REMARKS

TOC has been developed for more than 20 years. From the creation of theory and its improvement to the stage of practice, members of the TOC team have been committed to various researches on TOC, and have achieved some of the above-mentioned considerable results. Therefore, TOC has gradually become popular among people. In the field of vision, we firmly believe that the future of ternary optical computer is bright, but at the same time, we are also soberly aware that the course of its research and development will be arduous and tortuous. We hope that more people who are interested in ternary optical computer research will join our team and work together for the future of ternary optical computer.

REFERENCE

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

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

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

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

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

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

[7] Jin Yi, Ouyang Shan, Song Kai, Shen Yunfu, Peng Junjie, Liu Xuemin. Data Bit Management Theory and Technology of Ternary Optical Processor [J]. China Science (Information Science), 2013, 43(03):361 -373.

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

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

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

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

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

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

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

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

One Novel Soft-Starting Control Strategy for Induction Motor Based on Space Voltage Vectors

Shihong Xie
School of Electrical and Control Engineering
Shaanxi University of Science and Technology
Xi'an China
E-mail: skd2022xin@163.com

Zhihao Yang
School of Electrical and Control Engineering
Shaanxi University of Science and Technology
Xi'an China
E-mail: 1063496939 @qq.com

Rongmao Liang
School of Electrical and Control Engineering
Shaanxi University of Science and Technology
Xi'an China
E-mail: lrmtongxue@163.com

Xin Gao
School of Electrical and Control Engineering
Shaanxi University of Science and Technology
Xi'an China
E-mail: 945788415 @qq.com

Li Liang
School of Electrical and Control Engineering
Shaanxi University of Science and Technology
Xi'an China
E-mail: 3204958360 @qq.com

***Abstract*—This study presents a novel control strategy of discrete variable frequency (DVF) based on space voltage vectors for a three-phase induction motor (IM) soft-starting. The proposed strategy is developed in four steps. First, conducting states of the three-phase thyristor circuits are analysed, and the formation mechanism of space voltage vectors is also demonstrated. Then, DVF control strategy of IM based on hexagon flux linkage loci is expounded categorically, which contains frequency f/7 control, frequency f/4 control, frequency f/3 control and ramp voltage control. Next, under two-phase stationary coordinate system, stator flux equation of IM is derived, and the decrease in stator flux when zero-vector voltage working is calculated accurately. In the end, an experimental system of IM based on three-phase thyristor circuits is constructed, and the experimental results are dealt with and compared with the usual ramp voltage technique. The study results show that IM driven by the proposed strategy can steadily operate at the preset discrete frequency, and the effective value of starting current of IM can be decreased by nineteen percent. The experimental results verify the conclusions.**

Keywords-Induction Motor; Discrete Variable Frequency (DVF); High Starting Torque; Space Voltage Vectors; Thyristor

I. INTRODUCTION

Three-phase induction motors (IMs) are widely utilized in modern industry such as water pumps, draught fans, grinding millers and so on. Large starting currents will damage IM if it is directly started with power supply. So three-phase anti-parallel thyristor circuits based on voltage-regulation technology are often adopted for the soft start of IM [1], which have small starting current. However, using this way, electromagnetic torque of IM has more sacrifice than the decline of its starting current. Though many papers have proposed several methods to enhance start torque of IM, but they are based on constant-current control or closed-torque control [2-3]. [4] and [5] propose a novel control strategy of soft start for IM using pulse width modulation(PWM) AC

DOI: 10.2478/ijanmc-2022-0037

chopper, which uses four insulated gate bipolar transistors(IGBT) to regulate three-phase voltages of IM. This control method is simple and flexible, but the frequency of the three-phase voltages maintains a constant value. The promotion of IM starting torque in papers [2-5] is limited, because these methods all belong to the voltage-regulation control theory of IM.

Based on the three-phase thyristor circuits, Ginart and his cooperators propose a discrete variable frequency (DVF) theory, which can enhance electromagnetic torque of IM with the same starting current [6]. So, DVF theory is welcomed and further investigated in worldwide. In [7], phase control method of DVF optimal switching is studied, which has a favorable effect on solving switch disturbance but don't compare with other methods. In [8-9], one triggering scheme of output equivalent sinusoid voltages based on DVF is proposed, which analyzes the voltage symmetry and acquired available frequency dividing coefficients. It is insufficient that the control method can't consider the flux of IM when two-phase thyristor circuits are triggered. In [10-11], the cause of torque pulsation is studied and the proposed DVF control strategy is based on space voltage vectors, which can decrease the torque pulsation. However, the working principle of space voltage vectors and stator flux of IM isn't analyzed adequately in these papers. In [12] and [13], the causes of electromagnetic torque shocking and rotor speed shocking are analyzed with torque functions and simulation, and the technique used to suppress electromagnetic torque shocking is based on the closed control strategy of power factor angle compensating. While, there are large harmonic components and torque pulses for these DVF control methods proposed in above papers, which are based on periodic wave control theory. So torque increments of these methods are limited. There are other control methods in [14]-[16] based on AC chopper technology for soft-starting of IM, but they are mainly based on regulation voltage theory which can't increase starting torque of IM thoroughly.

In context to insufficient information existing in above the papers, based on [10-11], this paper proposes one novel DVF control strategy of IM based on space voltage vectors, which is realized with three-phase thyristor circuits. This novel control strategy is built on hexagon space voltage vectors and the stator flux linkage loci are controlled directly. Because the frequency of three-phase voltages is reduced with the decline of the effective value of the voltages. So the stator flux and the starting torque of IM driven by the proposed strategy are larger than the ramp voltage control.

II. DVF Principle Based on Space voltage vectors

A. Space voltage vectors based on three-phase thyristor control circuits

According to the principle of space voltage vectors, stator voltage vectors of IM can be written as follows [17].

$$\boldsymbol{u}_s = \sqrt{\frac{2}{3}}(u_{A0}e^{j0} + u_{B0}e^{j2\pi/3} + u_{C0}e^{j4\pi/3}) \qquad (1)$$

Where u_{A0}, u_{B0} and u_{C0} are three-phase stator winding voltages, respectively, and e^{j0}, $e^{j2\pi/3}$ and $e^{j4\pi/3}$ are unit space vector at the directions of A-phase, B-phase and C-phase stator winding axis of IM, respectively.

Thyristor control circuits of IM have three working states, which are two-phase circuits conducting, three-phase circuits conducting and three-phase circuits non-conducting. Space voltage vector will generate when two-phase windings of the IM are supplied with a power source, which are shown as fig.1.

Because stator current of IM will generate when at least two-phase circuits of thyristor control circuits are triggered. So when T1 and T6 are triggered, voltage space vector defined as $\boldsymbol{u}_{AB}$ which is shown in fig. 1(a) will generate in stator winding space of IM. When T1 and T2 are triggered, voltage space vector defined as $\boldsymbol{u}_{AC}$ in fig. 1(b) will generate. Similarly, $\boldsymbol{u}_{BC}$, $\boldsymbol{u}_{BA}$, $\boldsymbol{u}_{CA}$ and $\boldsymbol{u}_{CB}$ are presented in fig. 1(c) to (f). Expressions of following six space voltage vectors are shown here in equations (2) when the initial phase angle of phase-A voltage is equal to zero.

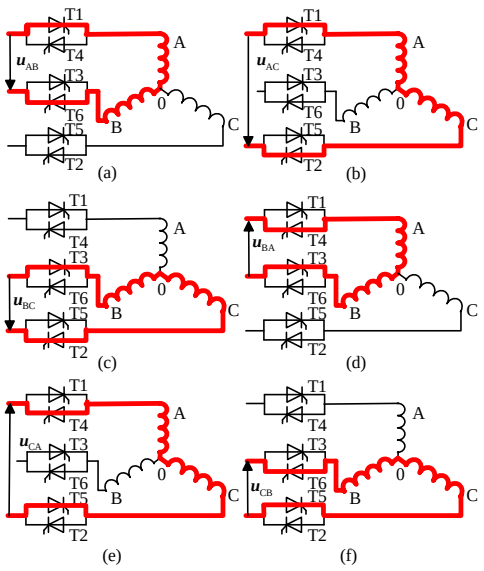


Figure 1. The forming principle of space voltage vectors based on three-phase thyristor circuits. (a) uAB, (b) uAC, (c) uBC, (d) uBA, (e) uCA, (f) uCB.

$$\begin{cases} \boldsymbol{u}_{AB}=\sqrt{\frac{2}{3}}U_m\left(\cos\omega_1 t-\cos(\omega_1 t-\frac{2\pi}{3})e^{\frac{j2\pi}{3}}\right) \\ \boldsymbol{u}_{AC}=\sqrt{\frac{2}{3}}U_m\left(\cos\omega_1 t-\cos(\omega_1 t+\frac{2\pi}{3})\bullet e^{\frac{j4\pi}{3}}\right) \\ \boldsymbol{u}_{BC}=\sqrt{\frac{2}{3}}U_m\left(\cos(\omega_1 t-\frac{2\pi}{3})e^{\frac{j2\pi}{3}}-\cos(\omega_1 t+\frac{2\pi}{3})e^{\frac{j4\pi}{3}}\right) \\ \boldsymbol{u}_{BA}=\sqrt{\frac{2}{3}}U_m\left(\cos(\omega_1 t-\frac{2\pi}{3})e^{\frac{j2\pi}{3}}-\cos(\omega_1 t)\right) \\ \boldsymbol{u}_{CA}=\sqrt{\frac{2}{3}}U_m\left(\cos(\omega_1 t+\frac{2\pi}{3})e^{\frac{j4\pi}{3}}-\cos(\omega_1 t)\right) \\ \boldsymbol{u}_{CB}=\sqrt{\frac{2}{3}}U_m\left(\cos(\omega_1 t+\frac{2\pi}{3})e^{\frac{j4\pi}{3}}-\cos(\omega_1 t-\frac{2\pi}{3})e^{\frac{j2\pi}{3}}\right) \end{cases} \quad (2)$$

Where Um is the peak value of phase voltage, ω1 is the angular frequency of power sources. Meanwhile, the voltage vector that three-phase thyristor circuits are all triggered is defined as $\boldsymbol{u}_{ABC}$. Similarly, the voltage vector with that three-phase thyristor circuits are all closed is defined as $\boldsymbol{u}_0$. According to the actual conducting circuits, $\boldsymbol{u}_{ABC}$ can be compounded with two space voltage vectors in (2).

If the voltage of stator resistance of IM is ignored, the stator flux of IM can be shown as the following [18].

$$\psi=\int(u+R_s\bullet i)dt\approx\int udt \quad (3)$$

Where ψ is the stator flux of IM, u is the stator voltage and R_s is the stator resistance. If the six space voltage vectors shown in (2) are connected in sequence, then one hexagon space voltage vectors will generate and stator flux linkage loci are also one hexagon as in fig.2 shown.

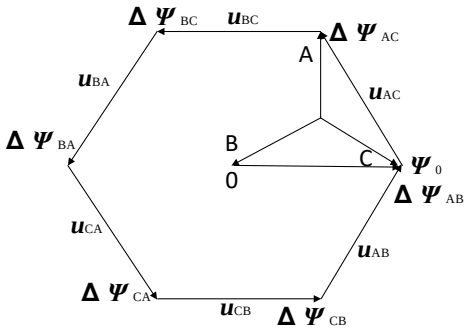


Figure 2. Hexagon space voltage vectors

B. DVF control principle based on space voltage vectors

Depending on the demand of voltage symmetry, the coefficient dividing the frequency of power voltage should be one, four and seven and so on. In case of starting torque of IM, the frequency f/7 can be chosen as the starting frequency of a heavy load IM. Although, the frequency f/3 of voltage doesn't meet the symmetry demand, but its third-harmonic component is a power frequency voltage. So the frequency f/3 of voltage has a good effect on the IM in fact. Therefore, voltage with frequency f/3 can also be utilized to drive an IM. Hence frequency division coefficient of power voltage of DVF would be seven, four, three, and one.

Control methods of DVF frequency f/7 based on space voltage vectors can be realized in following way:

First, based on the zero passage of rising edge of phase-A voltage, T1 and T2 are triggered in order to generate $\boldsymbol{u}_{AC}$, which will trigger angle θ_1. When current i_{AC} declines to zero, T1 and T2 will be closed naturally. Then, during next primitive period of power source, based on the zero passage of falling edge of phase-C voltage, T3 and T2 are triggered in order to generate $\boldsymbol{u}_{BC}$. Their trigger

angles are all θ_2. Circulating this way, after thyristors are closed every time, next vector of hexagonal space voltage vectors will be generated during the next primitive period of the power source. All of hexagonal voltage vectors are generated once during seven primitive periods of the power source. So voltage with frequency f/7 has waves of DVF based on hexagon space voltage vectors is acquired as fig.3 shown.

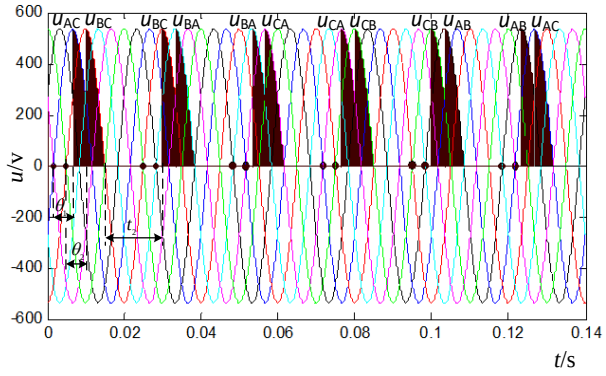


Figure 3. Frequency f/7 voltage waves of DVF based on space voltage vectors

Voltages with frequency f/4 based on three-phase power frequency sinusoid voltages are positive voltages. And control methods of DVF frequency f/4 are also based on hexagonal stator flux linkage loci. But the difference with the control methods of DVF frequency f/7 is that number of working vectors of the former is half of the latter. So voltage with frequency f/4 can be acquired by selectively reducing voltage vectors of frequency f/7 voltage vectors. The sequences of effective voltage vectors are $\boldsymbol{u}_{AC}$ to $\boldsymbol{u}_{BC}$ to $\boldsymbol{u}_{BA}$ to $\boldsymbol{u}_{CA}$ to $\boldsymbol{u}_{CB}$ to $\boldsymbol{u}_{AB}$. The stator flux linkage loci will be three continuous sections which consist of $\boldsymbol{\psi}_{AC}$, $\boldsymbol{\psi}_{BC}$, $\boldsymbol{\psi}_{BA}$, $\boldsymbol{\psi}_{CA}$, $\boldsymbol{\psi}_{CB}$, and $\boldsymbol{\psi}_{AB}$. Frequency f/4 Voltage waveforms are shown in fig.4.

Sequences of effective space voltage vectors of frequency f/3 are $\boldsymbol{u}_{AC}$, $\boldsymbol{u}_{BC}$, $\boldsymbol{u}_{BA}$, $\boldsymbol{u}_{CA}$, $\boldsymbol{u}_{CB}$, and $\boldsymbol{u}_{AB}$, which are also based on hexagonal stator flux linkage loci, but they are divided into two sets. The first set is continuously triggered, containing $\boldsymbol{u}_{AC}$, $\boldsymbol{u}_{BC}$, and $\boldsymbol{u}_{BA}$ while second set is also continuously triggered, which contain $\boldsymbol{u}_{CA}$, $\boldsymbol{u}_{CB}$, and $\boldsymbol{u}_{AB}$. Voltage waveforms of frequency f/3 are shown in fig.5.

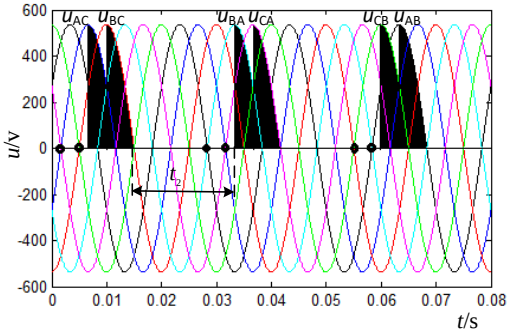


Figure 4. Frequency f/4 voltage waves of DVF based on space voltage vectors

Based on above control methods and space voltage vectors, soft start control of IM is operated in following methods.

Frequency f/7 is selected as the starting frequency of IM. Before soft start, IM is be excited with one voltage vector in several power frequency periods. In order to acquire suitable stator flux, the pre-excited voltage vector will be the previous of the initial voltage vector of frequency f/7, according to the order of hexagon space voltage vectors. For example, if $\boldsymbol{u}_{AC}$ is the first working voltage vector of frequency f/7, then $\boldsymbol{u}_{AB}$ will be the pre-excitation voltage vector. Time for pre-excitation can be calculated by (3) according to the actual IM parameters.

After pre-excitation, IM is driven with the method of DVF having frequency f/7, then frequency f/4, frequency f/3 and frequency f/1 as shown in fig. 3 to 5. Frequency f/1 is the traditional ramp voltage control.

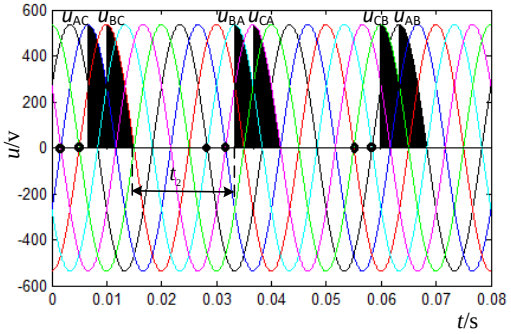


Figure 5. Frequency f/3 voltage waves of DVF based on space voltage vectors

When the working frequency of IM is changed, the previous voltage vector and the next voltage

space vector accord with the sequence of hexagon space voltage vectors as shown in fig. 2. For example, when the working frequency changes from frequency f/7 to frequency f/4, so if $\boldsymbol{u}_{BC}$ is the final voltage vector of frequency f/7 space voltage vectors, then $\boldsymbol{u}_{BA}$ will be the first voltage vector of frequency f/4 voltage vectors. Similarly, when working frequency changes from frequency f/4 to frequency f/3, so if $\boldsymbol{u}_{AB}$ is the final voltage vector of frequency f/4 voltages, then $\boldsymbol{u}_{AC}$ will be the first voltage vector of frequency f/3 voltage vectors. After frequency f/3, IM is controlled by the ramp voltage control.

C. Stator flux analysis of IM

Space voltage vectors of frequency f/3 for IM is used for the stator flux analysis. Stator windings of IM controlled by three-phase thyristor circuits are shown in fig.6. When initial conducting circuits are phase-A and phase-B, and stator windings of IM are star connection, then the current in phase-C stator winding is equal to zero. So phase-C stator winding can be located on α axis of the α-β stationary reference frame, and the equations are acquired as the following.

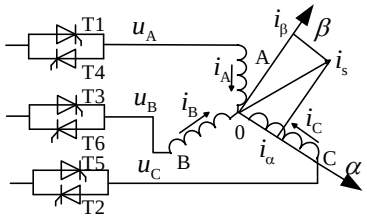


Figure 6. Circuit model of IM under α-β stationary reference frame

$$i_{s\alpha} = 0\ , \quad i_A = -i_B \tag{4}$$

$$i_{s\beta} = \sqrt{\frac{2}{3}}\left(\frac{\sqrt{3}}{2}i_B - \frac{\sqrt{3}}{2}i_A\right) = -\sqrt{2}i_A \tag{5}$$

$$u_{s\beta} = \sqrt{\frac{2}{3}}\left(\frac{\sqrt{3}}{2}u_A - \frac{\sqrt{3}}{2}u_B\right) = \sqrt{2}u_{AB} \tag{6}$$

Where $i_{s\alpha}$ and $i_{s\beta}$ are the α-axis stator current and the β-axis stator current respectively under α-β stationary reference frame, and $u_{s\beta}$ is the stator voltage of α-β stationary reference frame. i_A and i_B are respectively the current in phase-A and phase-B of IM under A-B-C three-phase stationary reference frame. u_{AB} is the line voltage of phase-A to phase-B stator windings. u_{AB} is shown as the following.

$$u_{AB}(t) = \sqrt{3}U_m \sin(\omega_1 t + \alpha_0) \tag{7}$$

Where α_0 is trigger angle of T1 and T6.

Based on the model of IM and from (4) to (7) equations, the following equations are acquired by using Laplace transform.

$$\begin{cases} U_{AB}(s) = -(R_s + sL_s)I_A(s) + \sqrt{2}sL_m I_{r\beta}(s)/2 \\ 0 = (R_r + sL_r)I_{r\alpha}(s) + \omega_r L_r I_{r\beta}(s) - \sqrt{2}\omega_r L_m I_A(s) \\ 0 = (R_r + sL_r)I_{r\beta}(s) - \sqrt{2}sL_m I_A(s) - \omega_r L_r I_{r\alpha}(s) \end{cases} \tag{8}$$

Where L_m is mutual inductance between stator windings and rotor windings, L_s is stator self-inductance, L_r is rotor self-inductance, R_s is stator resistance, R_r is rotor resistance, $I_A(s)$ is the current in phase-A stator winding. $I_{r\alpha}(s)$ and $I_{r\beta}(s)$ are the α axis current and the β axis current of IM under α-β stationary reference frame respectively. ω_r is angular frequency of IM. In order to simplify the progress of solving $I_A(s)$, ω_r is set to be zero. Then IA(s) is solved by using (8).

$$\begin{cases} I_A(s) = P(s)/Q(s) \\ P(s) = \dfrac{\sqrt{3}U_m(-s\cos\alpha_0 + \omega_1 \sin\alpha_0)(1+sT_r)}{R_s T_s T_r'} \\ Q(s) = \left(s^2 + \omega_1^2\right)\left(s^2 + s(\dfrac{1}{T_r'} + \dfrac{1}{T_s'}) + \dfrac{1}{T_s' \bullet T_r'}\right) \end{cases} \tag{9}$$

Where σ=1-L^2_m/(L_sL_r), T_s=L_s/R_s, T_r=L_r/R_r, T'_s=σT_s、T'_r=σT_r. When $Q(s)$ is equal to zero, then there are four roots that can be acquired from the equation. Two of the four roots are complex conjugate: $\lambda_{1,2}$=±jω_1, which correspond to steady components of $I_A(s)$. Another two roots are signed as λ_3 and λ_4. The real parts of λ_3 and λ_4 are negative, which correspond to transient

components of $I_A(s)$. $I_{r\beta}(s)$ are also acquired by using (8).

$$I_{r\beta}(s)=\frac{\sqrt{6}U_m(s\cos\alpha_0-\omega_1\sin\alpha_0)sL_m}{L_sL_r\sigma\left(s^2+\omega_1^2\right)\left(s^2+s(\frac{1}{T_r'}+\frac{1}{T_s'})+\frac{1}{T_s'\bullet T_r'}\right)} \tag{10}$$

Based on the stator flux model of IM: $\boldsymbol{\psi}_s=L_s\boldsymbol{i}_s+L_m\boldsymbol{i}_r$, $\boldsymbol{\psi}_s$ can be calculated by using (4), (5), (8), (9) and (10),

$$\begin{aligned}\boldsymbol{\psi}_s(s)&=j(L_mI_{r\beta}(s)-\sqrt{2}L_sI_A(s))\\&=j\frac{\sqrt{6}U_m(s\cos\alpha_0-\omega_1\sin\alpha_0)\left((1-\sigma+\frac{1}{L_s})s+\frac{1}{L_sT_r}\right)}{\sigma\left(s^2+\omega_1^2\right)\left(s^2+s(\frac{1}{T_r'}+\frac{1}{T_s'})+\frac{1}{T_s'\bullet T_r'}\right)}\end{aligned} \tag{11}$$

Equation (11) can be rewritten by using Laplace inverse transformation as the following.

$$\begin{aligned}\boldsymbol{\psi}_s(t)&=L^{-1}\psi_s(s)\\&=K_1e^{-j\omega_1t}+K_2e^{j\omega_1t}+K_3e^{\lambda_3t}+K_4e^{\lambda_4t}\end{aligned} \tag{12}$$

Where

$$\begin{cases}K_1=\dfrac{\sqrt{6}U_m\omega_1(-\cos\alpha_0+j\sin\alpha_0)\left(-\left(1-\sigma+\frac{1}{L_s}\right)j\omega_1+\frac{1}{L_sT_r}\right)}{\sigma(2j\omega_1)(j\omega_1+\lambda_3)(j\omega_1+\lambda_4)}\\K_2=\dfrac{\sqrt{6}U_m\omega_1(-\cos\alpha_0-j\sin\alpha_0)\left(\left(1-\sigma+\frac{1}{L_s}\right)j\omega_1+\frac{1}{L_sT_r}\right)}{\sigma(2j\omega_1)(j\omega_1-\lambda_3)(j\omega_1-\lambda_4)}\\K_3=\dfrac{j\sqrt{6}U_m(\lambda_3\cos\alpha_0-\omega_1\sin\alpha_0)\left(\left(1-\sigma+\frac{1}{L_s}\right)\lambda_3+\frac{1}{L_sT_r}\right)}{\sigma\left(\lambda_3^2+\omega_1^2\right)(\lambda_3-\lambda_4)}\\K_4=\dfrac{j\sqrt{6}U_m(\lambda_4\cos\alpha_0-\omega_1\sin\alpha_0)\left(\left(1-\sigma+\frac{1}{L_s}\right)\lambda_4+\frac{1}{L_sT_r}\right)}{\sigma\left(\lambda_4^2+\omega_1^2\right)(\lambda_4-\lambda_3)}\end{cases} \tag{13}$$

The stator flux function is acquired by the same method, when $\boldsymbol{u}_{AC}$ and $\boldsymbol{u}_{BC}$ are working. Similarly, $\boldsymbol{u}_0$ is working, then stator current of IM is equal to zero, and the rotor current of IM is shown as following.

$$i_r(t)=i_{r0}e^{-t/T_r} \tag{14}$$

When $\boldsymbol{u}_0$ is working, fig.5 shows that max value of t_3 is 23.3ms and minimum value of t_3 is 13.3ms. When trigger angle becomes equal to ninety degrees, t_3 is 16.7ms. For a common fifteen kilowatt IM, rotor time constant is approximately equal to 300ms which is much larger than t_3. So, based on these parameters and the stator flux model of IM: $\boldsymbol{\psi}_s=L_s\boldsymbol{i}_s+L_m\boldsymbol{i}_r$, stator flux ψ_s can be calculated by using the following equation.

$$\psi_s(t)=L_mi_{r0}e^{-t/T_r}=\psi_0e^{-t/T_r}=0.955\psi_0 \tag{15}$$

Where, ψ_0 is the initial value of ψ_s. Equation (15) shows that the decrement of stator flux is very small and can be neglected when $\boldsymbol{u}_0$ is working.

Using the same method, stator flux can be calculated when IM is also driven by frequency f/7 control method or frequency f/4 control method of DVF.

III. Experiment Validation

A. Simulation

The simulation model is built by Matlab software. Parameters of IM used in the model are shown as the following: P_N=15kW, U_N=380V, n_N=1460r/min, I_N=29.5A, f_N=50Hz, Lm=64.19mH, Rs=0.2147Ω, Rr=0.2205Ω, J=0.602 kg·m^2, $L_{s\sigma}=L_{r\sigma}$=0.991mH. Load rate of IM is 60 percent. Simulation results are shown in fig. 7 to 12.

When IM is driven by control method of DVF with frequency f/7, its line voltage and phase current are shown in fig. 7. The simulation results in fig. 7 show that the period of the line voltage is 0.14ms, and phase current has four continuous conducting sections in time of 0.14ms. So the simulation results in fig.7 verify the principle of control method of DVF with frequency f/7.

Line voltage and phase current of the IM driven by control method of DVF with frequency f/4 are shown in fig 8(a) and fig 8(b) respectively. Fig 8(a) shows that the period of the line voltage is 0.08ms and fig 8(b) shows that phase current has three continuous conducting sections in one frequency f/4 period. The simulation results in fig.8 verify the principle of control method of DVF with frequency f/4.

Similarly, line voltage and phase current for frequency f/3 of DVF are shown in fig. 9. Fig. 9(a) shows that the period of the line voltage is 0.06ms and fig. 9(b) shows that phase current has two continuous conducting sections in time of 0.06ms. The simulation results in fig.9 verify the principle of frequency f/3 control method of DVF.

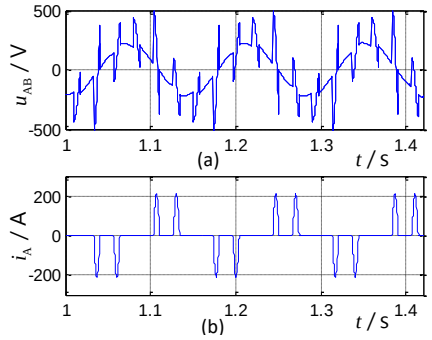


Figure 7. Voltage and current for frequency f/7. (a)Line Voltage. (b) Phase Current.

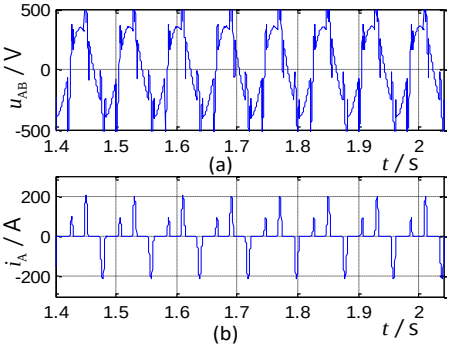


Figure 8. Voltage and current for frequency f/4. (a) Line Voltage. (b) Phase Current.

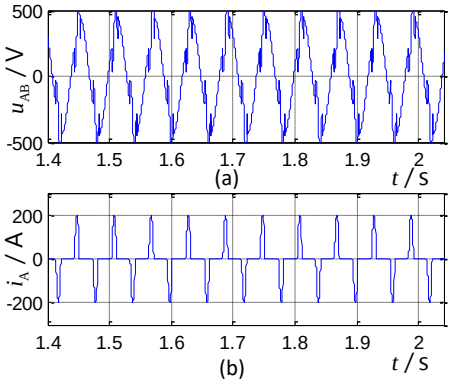


Figure 9. Voltage and current for frequency f/3. (a) Line Voltage. (b) Phase Current.

Stator flux of IM driven by the proposed strategy and the ramp voltage control are shown in fig.10. The results show that the stator flux amplitude of IM driven by the proposed strategy is larger than the flux under the traditional ramp voltage control. The reason is that the frequency of voltage of DVF decreases, but the frequency of voltage of ramp voltage control is a constant value, in the process of soft starting of IM.

Stator current and rotor speed of IM based on the proposed strategy are shown in fig. 11. The result in fig. 11(a) shows that the currents includes four sections which successively corresponded to frequency f/7 current, frequency f/4 current, frequency f/3 current and ramp voltage regulation current. Meanwhile, rotor speed of IM, shown in fig. 11(b), increases accordingly to the proposed control strategy. The simulation results verify the principle of space voltage vectors control of DVF.

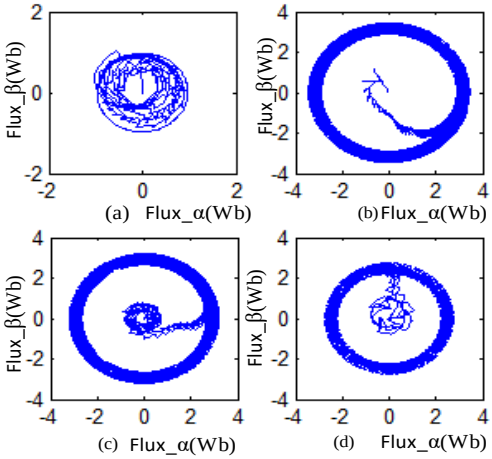


Figure 10. Induction motor stator flux track. (a) ramp voltage. (b) frequency f/7 (c) frequency f/4. (d) frequency f/3.

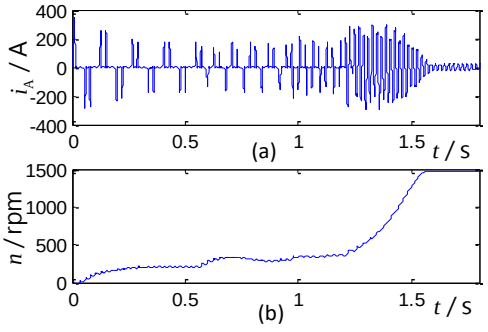


Figure 11. Responses of IM driven by the proposed method. (a) Stator current. (b) Rotor speed.

Contrary to this, stator current and rotor speed of IM driven by ramp voltage control are shown in fig. 12. The results show that rotor speed increases quickly and the value of starting current is very large, which all conform to the principle of ramp voltage soft-starting of IM.

Comparing fig.11 and 12, the current in fig. 11 is smaller than the current in fig. 12, and the rotor speed increases softly. So, the proposed control strategy can acquire better starting performance than ramp voltage control for IM.

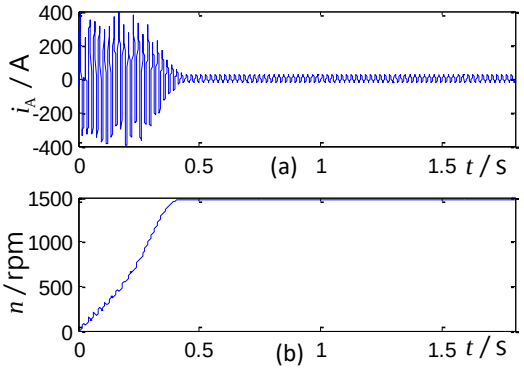


Figure 12. Responses of IM driven by the traditional method. (a) Stator current. (b) Rotor speed.

B. Experiment

For further verifying the performance of the proposed strategy, one experimental set of IM based on three-phase thyristor circuits and STM32F103RC microcontroller is designed as shown in fig. 13. The parameters of IM are same to the parameters used in the simulation model. The experimental results are shown in figs. 14-17.

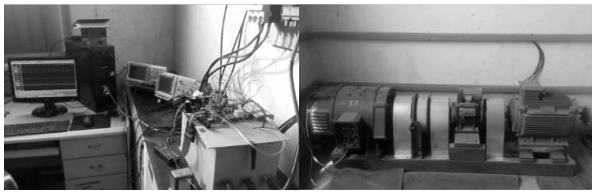


Figure 13. Experimental system of IM driven by three-phase thyristor circuits

Fig. 14 shows the experimental results of IM driven by frequency f/7 of the proposed control strategy. Fig. 14(a) shows the voltage of phase-A to phase-B, while fig. 14(b) shows the current in phase A of IM. They correspond to fig.7 (a) and fig. 7(b) respectively. The experimental results show that the period of the voltage is 0.14ms, and the current has four conducting sections in the period. The experimental results in fig. 14 are in accordance with the simulation results in fig. 7.

Fig. 15 shows the experimental results of IM driven by frequency f/4 of the proposed control strategy. The experimental results in fig. 15 correspond to the simulation results in fig.8. Fig. 15(a) is the voltage of phase-A to phase-B and fig. 15(b) is the current in phase-A of IM. The experimental results show that the period of the voltage is 0.08ms, and the current has three conducting sections in the period. The experimental results in fig.15 are in accordance with the simulation results in fig.8.

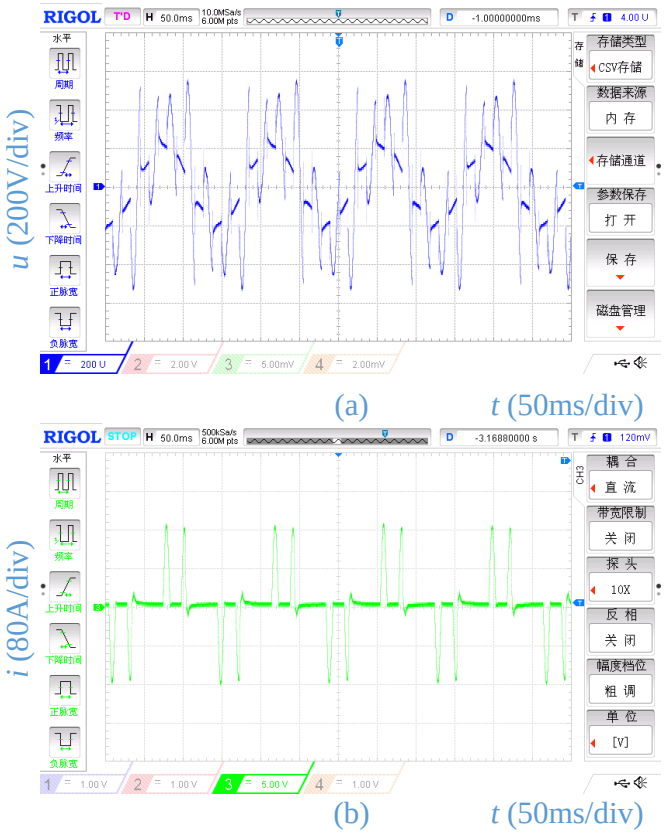


Figure 14. Voltage and current for frequency f/7. (a) Line Voltage (b) Phase Current.

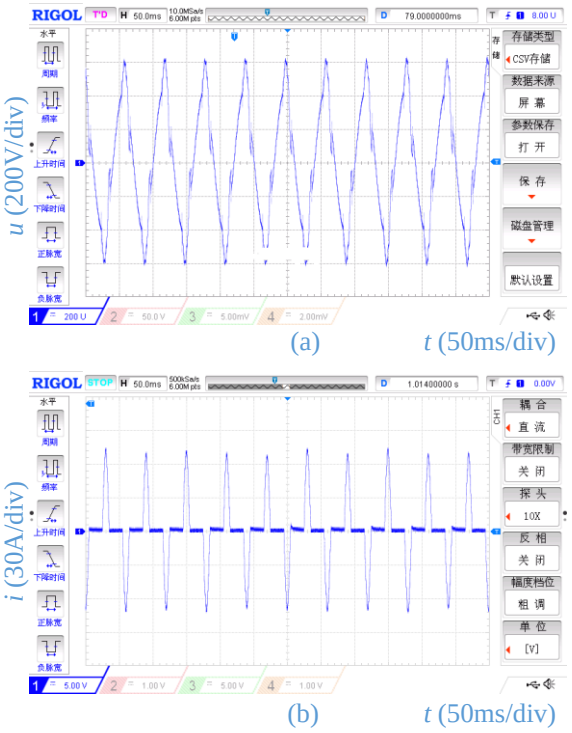


Figure 15. Voltage and current for frequency f/4. (a) Line Voltage. (b) Phase Current.

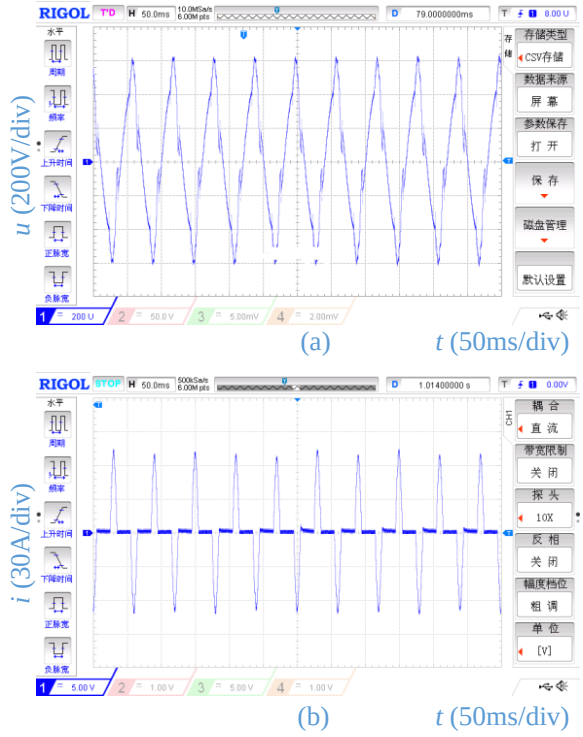


Figure 16. Voltage and current for frequency f/3. (a) Voltage waveform. (b) Current waveform.

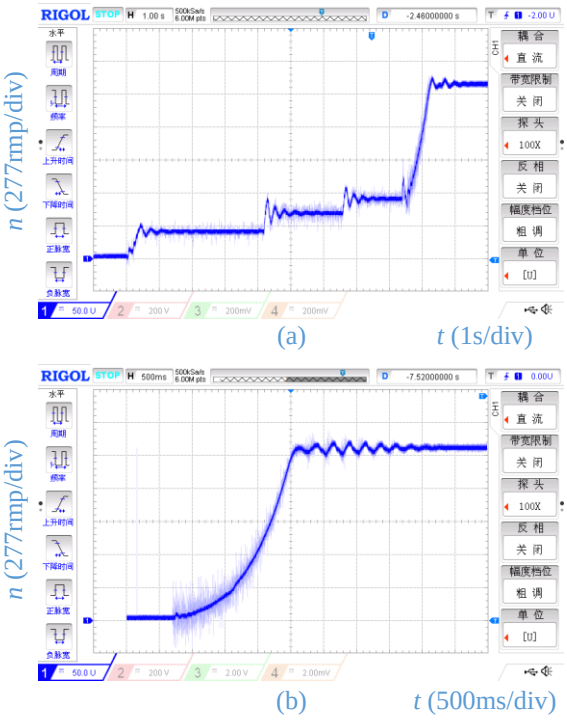


Figure 17. Rotor speeds of IM. (a) The proposed method. (b) The traditional method

Fig.16 shows the experimental results of IM driven by frequency f/3 of the proposed control strategy. The experi-mental results in fig. 16 correspond to the simulation results in fig. 9. Fig. 16(a) is the voltage of phase-A to phase-B and the fig.16 (b) is the current in phase-A of IM. The experimental results show that the period of the voltage is 0. 06ms, and the current has two conducting sections in the period. The experimental results in fig. 16 are in accordance with the simulation results in fig. 9.

Fig.17 demonstrates the rotor speeds of IM driven by the proposed strategy and the traditional ramp voltage. Fig. 17 (a) shows that the rotor speed responds to the proposed strategy. Fig.17 (b) also shows that rotor speed responds to the ramp voltage control. Meanwhile, fig. 17(a) corresponds to the simulation result in fig.11 (b), and fig. 17(b) corresponds to the simulation result in fig.12 (b).

Another experimental results show that the effective value of starting current of IM driven by the proposed strategy with sixty percent load is one hundred and twenty four amperes (124 amps). Whereas when the motor is driven by the ramp voltage control with the same load, the effective

value of starting current is one hundred and fifty four amperes (154 amps). The starting current with the proposed strategy decreases by nineteen percent comparing with the traditional strategy.

IV. CONCLUSION

This paper studies the principle of space voltage vectors based on three-phase thyristor circuits, and proposes the control strategy of DVF based on hexagon space voltage vectors. The study results show that frequency f/7, frequency f/4 and frequency f/3 of power sources can be used to drive IM. The experimental results also show that starting current of a fifteen kilowatt IM driven by the proposed strategy decreases by nineteen percent as comparing to the traditional ramp voltage control with the same load, and rotor speed accurately changes according to the proposed strategy. So, the proposed control strategy of DVF based space voltage vectors is effective for soft starting of IM.

REFERENCES

[1] T.Veera, K.Vijit,K.Anantawat, Comparison of Starting Current Characteristics for Three-Phase nduction Motor Due to Phase-control SoftStarter and Asynchronous PWM AC Chopper," *Journal of Electronical Engineering & Technology*, 2017, Vol.12, No.3, pp. 1090-1100, 2017.

[2] X.Y. Li, J. Xu, H.P, Zhang. "Research on torque ramp current limit starting of induction motor based on dsPIC30F6014," *In Proc. of IEEE Inf. Tech. Net. Elec. Auto. Con.*, pp.1627-1630, Jan. 2018.

[3] A. Nied, J. de Oliveira, R. de F. Campos, et al, "Soft Starting of Induction Motor With Torque Control," *IEEE Trans. Ind. App.*, Vol.46,No.3, pp.1002-1010, 2010.

[4] Said A.D., Haitham Z.A., "Current limiting soft starter for three phase induction motor drive system using PWM AC chopper," IET Power Electronics,Vol.10, No.11, 1298-1306,2017.

[5] Khan, M.M., Rana, A., D., Fei. "Improved ac/ac choppers-based voltage regulator designs," IET Power Electronics, Vol.7, NO. 8, pp.1989-2000, 2014.

[6] A. Ginart, R. Esteller, A Maduro, et al. "High starting torque for AC SCR controller," IEEE Trans Energy Conversion, Vol.14, No.3, pp.553-559, 1999.

[7] F. Zhou, J.L. Cao, J. Liu, et al. "Optimal switching phase and frequency splitting strategy of discrete frequency conversion soft starting control for asynchronous motor," *Electric Machines and Control*, Vol.20,No.3, pp.13-19, 2016.

[8] S.J. Deng, J. Chen, Q. Wang, et al. "A controller of Motor Discrete Variable Frequency Soft Starting and Harmonic Filtering," *2nd Int. Conf. on Sensors, Measurement and Intelligent Materials*, pp.1611-1614, Dec. 2013.

[9] D.H., Li., X.B., Deng. "Research on Discrete Variable Frequency soft starting and electricity-economizing control system of induction motor,"*Int. Conf. Electr. Inf. Control Eng.*, pp. 4391-4394, 2011.

[10] J. Chen, Y.J. Meng, M.L. Duan, et al. "A Starting Method of Induction Motor with High Starting Torque," *Transaction of China Electrotechnical Society*, Vol.32, No.9, pp. 32-39, 2017.

[11] S.H. Xie, Y.J. Meng, J. Chen, et al. "Cause and improvement of the discrete variable frequency torque ripple of induction motor," *Electric Machines and Control*, Vol.22,No.10, pp. 103-111, 2018,.

[12] J. Tong, Z. Zhang, C.Y. Guo, "Study of power factor angle closed-loop control technology in soft-starter," *Electric Machines and Control*, 17, (12), pp. 51-56, 2013.

[13] K.Q. Zhao, D.G. Xu, Y. Wang, "New strategy to improve electromagnetic torque at starting in thyristor controlled induction motors," *The 29th Annual Conference of the IEEE Industrial Electronics Society*, pp.2555-2560, Nov. 2003.

[14] R. L. Gorbunov, G. I. Poskonnyy, "Symmetrical discrete frequency control for AC-chopper with mutual switching function," *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM*, pp. 353-358, Jul. 2014.

[15] K Sundareswaran, P.S R Nayak, "Particle Swarm Optimisation Based Feedback Controller Design for Induction Motor Soft-Starting," *Australian Journal of Electrical and Electronics Engineering*, Vol.11, No.1, pp. 55-63, 2014.

[16] P. Srinivasa Rao Nayak, T.A. Rufzal, "Performance analysis of feedback controller design for induction motor soft-starting using bio-inspired algorithms," *Proc. IEEE Int. Conf. Power, Instrum., Control Comput., PICC*, pp. 1-6, Jun. 2018.

[17] Q.T. An, F. Yao, L.Z. Sun, L. Sun, "SVPWM Strategy of Dual Inverters and Zero-Sequence Voltage Suppression Method," *Proceedings of the CSEE*, Vol.36, No.4, pp. 1042-1049, 2016.

[18] H.Q. Zhang, X.S. Wang, P.F. Wang, et al, "Study on direct torque control algorithm based on space vector modulation," *Electric Machines and Control*, Vol.16, No.6, pp. 13-18, 2012.

Super-resolution Reconstruction Based on Capsule Generative Adversarial Network

Ziyi Wu
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
xi'an, Shaanxi, China
E-mail: wuziyi_5817@163.com

Hong Jiang
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
xi'an, Shaanxi, China
E-mail: 249479898@qq.com

Hongge Yao
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
xi'an, Shaanxi, China
E-mail: yaohongge@xatu.edu.cn

Wei Zhang
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
Xi'an, Shaanxi, China
E-mail: 913917873@qq.com

Hualong Yang
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
xi'an, Shaanxi, China
E-mail: 1067726236@qq.com

Jun Yu
School of Computer Science and Engineering
Xi'an Technological University
No.2 Xuefu Middle Road, Weiyang district,
xi'an, Shaanxi, China
E-mail: 763757335@qq.com

***Abstract*—Using each part of the image's spatial information to generate better local details of the image is a key problem that super-resolution reconstruction has been facing. At present, mainstream super-resolution reconstruction networks are all built based on convolutional neural networks (CNN). Some of these methods based on Generative Adversarial Networks (GAN) have good performance in high-frequency details and visual effects. However, because CNN lacks the necessary attention to local spatial information, the reconstruction method is prone to problems such as excessive image brightness and unnatural pixel regions in the image. Therefore, using the capsule network's excellent perception of hierarchical spatial information and local feature relationships, the author proposes a super-resolution reconstruction based on capsule network CSRGAN. The experiment's final result shows that compared with the pure convolution method RDN, the PSNR value of CSRGAN is increased by 0.14, which is closer to the original image.**

Keywords-Generative Adversarial Network; Capsule Network; Capsule Generative Adversarial Network; Capsule Discriminator; Super-resolution Reconstruction

I. INTRODUCTION

In computer image processing, obtaining more necessary information from low-resolution input images to achieve super-resolution reconstruction (SR) is always a critical issue. To solve this problem, it needs to use the low-resolution image's existing information to infer its own mapping of the high-resolution image. The essence of this mapping is an ill-conditioned reasoning problem,

so super-resolution reconstruction of low-resolution images is an arduous task. Artificial intelligence methods have been intensely studied in such fields as intelligent cognition technology [1], efficient image retrieval technology [2], blockchain technology [3], communication virtual mobile network resource management [4], machine motion detection analysis [5], fault identification [6], and so on, and have achieved exciting research results. Moreover, CNN [7] and reinforcement learning are the most widely used in these fields among the artificial intelligence methods and have achieved exciting research results. To obtain better reconstruction results, many scholars have proposed several super-resolution reconstruction methods based on deep learning. Among them, the deep network based on CNN has shown great promise in reconstruction accuracy [8].

He Kaiming and others first proposed a super-resolution reconstruction network SRCNN that uses a three-layer convolutional network to fit nonlinear mapping in 2014 and achieved excellent experimental results at that time [9]. This is also the first successful attempt of CNN in the field of super-resolution reconstruction. In 2016, Wenzhe Shi and others proposed the ESPCN framework [10], Compared with those methods that use interpolation to enlarge the image to the target size before doing the convolution operation, the sub-pixel convolution mentioned in the paper directly performs the upsampling process on the low-resolution image, which makes the ESPCN significantly reduced the calculation complexity and parameter amount [11]. The running speed is also improved, and it can even support real-time super-resolution reconstruction in the video. In 2017, SRGAN [12] proposed by Christian Ledig et al. was the first successful attempt to use the GAN [13] to achieve super-resolution reconstruction. It performed well in improving the realism of the picture, generated better high-frequency texture details, and had good generalization, and it performs well in a scene with 4x resolution magnification. The emergence of SRGAN proves that GAN has excellent potential in the field of super-resolution reconstruction [14]. Based on the research of super-resolution reconstruction of deep convolutional networks [15] and deep spatial feature transform [16,17]. In 2018, Yulun Zhang et al. proposed the RDN [18]. The RDB structure used in this framework can make full use of all convolutional layers' hierarchical information to achieve more efficient feature learning and finally obtain outstanding visual effects. In 2020, the dual regression scheme used by Dual Regression Networks proposed by Yong Guo et al. added additional constraints to the LR data [19]. This framework allows the network to learn not only the mapping of low-resolution images to high-resolution images, but also the inverse mapping of super-resolution images to low-resolution images. These better mapping combinations can better guide the network to reconstruct images.

However, these super-resolution networks simply based on CNN usually pay more attention to the target's overall features but ignore the relationship between spatial level information and local features. In super-resolution reconstruction, enhancing the perception ability of local feature spatial information is of great significance for the reconstruction of local details. This characteristic of CNN hinders a super-resolution reconstruction network's process to create closer to the real details through this information and further leads to local over-smoothness of the generated image and sometimes even unnatural texture.

In order to solve the above problems, we propose a new generation of confrontation super-resolution reconstruction model CSRGAN. CSRGAN takes advantage of the Capsule Network's ability [20,21] to infer the relationship between the various parts of the image and makes up for the lack of CNN in dealing with the feature space connection. The CSRGAN model uses the RDN architecture as a generator to enhance the acquisition of each convolutional layer's hierarchical information and uses the Capsule network to replace the CNN network as a discriminator to enhance the model's ability to judge spatial level details. And by adding a vector inner product loss function to improve the picture quality of the generated picture. CSRGAN mainly has the following two contributions:

1) A dual-route capsule network discriminator is constructed, and the parameter matrix of the

capsule network in the discriminator is restricted small parameters through a clip function. It enables the discriminator network to extract features from coarse-grained to fine-grained through a dual-routing process, so that the capsule network can improve the efficiency and stability and also enhance the perception of the local spatial posture information of the generated image.

2) The vector inner product loss is proposed and added to the generator's loss function to train the generator network. The feature set extracted by the capsule network is also added to the loss calculation so that the generation network has a better representation of the local feature content and ultimately improves the performance of the local detail texture.

This article is composed of six chapters. The first chapter INTRODUCTION puts forward the motivation and contribution of this article; The second chapter RELATED WORK introduces previous work related to this article and their contribution to this article; The third chapter METHOD introduces the method proposed in this article and the network operation mechanism in detail; Chapter 4 LOSS FUNCTION introduces the theoretical basis of the network; Chapter 5 EXPERIMENTS is the specific experimental process and experimental data analysis of this article; Chapter 6 CONCLUSION is a summary of the full text.

II. REALATED WORK

SRGAN The SRGAN model inherits the GAN model's generative confrontation idea, and on its basis, improves the cost function for super-resolution reconstruction. One part of the improvement is the content-based cost function, and the other part is the cost function based on confrontation learning. The content-based cost function contains a minimum mean square error based on the feature space, calculated using the high-level features of the image extracted by the VGG [22]. These improvements make SRGAN better at processing high-frequency texture details. In the generation network, SRGAN uses the sub-pixel convolution in ESPCN for up-sampling, which improves the quality of the generated pictures and reduces the complexity of reconstruction calculations.

to

RDN The RDN model combines the ideas of Res-Net [23] and Dense-Net [24] and improves on it. RDN realizes the local feature fusion and global feature fusion of all feature maps' levels through the combined use of the Residual dense blocks (RDB) module and Dense feature fusion (DFF), achieving the maximum reuse of convolutional network-level information. This enables RDN to perform a more detailed feature understanding of the input low-resolution images when performing super-resolution tasks. And it can maintain a stable training effect even when the network level is deepened.

Capsule Network Capsule Network is a new type of network proposed by the Hinton team in 2017. Hinton believes that although the "translation invariance" of CNN brings the robustness of classification, it inevitably loses the spatial hierarchical information of the feature, and this information is just helpful to the understanding of the feature. The capsule network uses vectors to express feature information. Its "identity" can represent the spatial hierarchical information in the feature, and it has better recognition in image classification from different angles. The modulus of the capsule vector represents the probability of the feature's existence, and the direction of the vector represents the spatial hierarchical relationship between the features. These characteristics of the capsule vector make the capsule network more accurate in feature recognition. Moreover, the capsule network can complete network training with less image training data.

Based on the advantages of the capsule network and the above two networks, we use RDN to build the basic structure of the generation network and use the capsule network to build the discriminator's basic structure. Besides, we have implemented dual routing improvements based on the capsule network. The two routing processes of the capsule cooperate so that the network has a better perception of the local correlation of the image feature space, and finally promotes the generator to generate a closer high-resolution image of the original image.

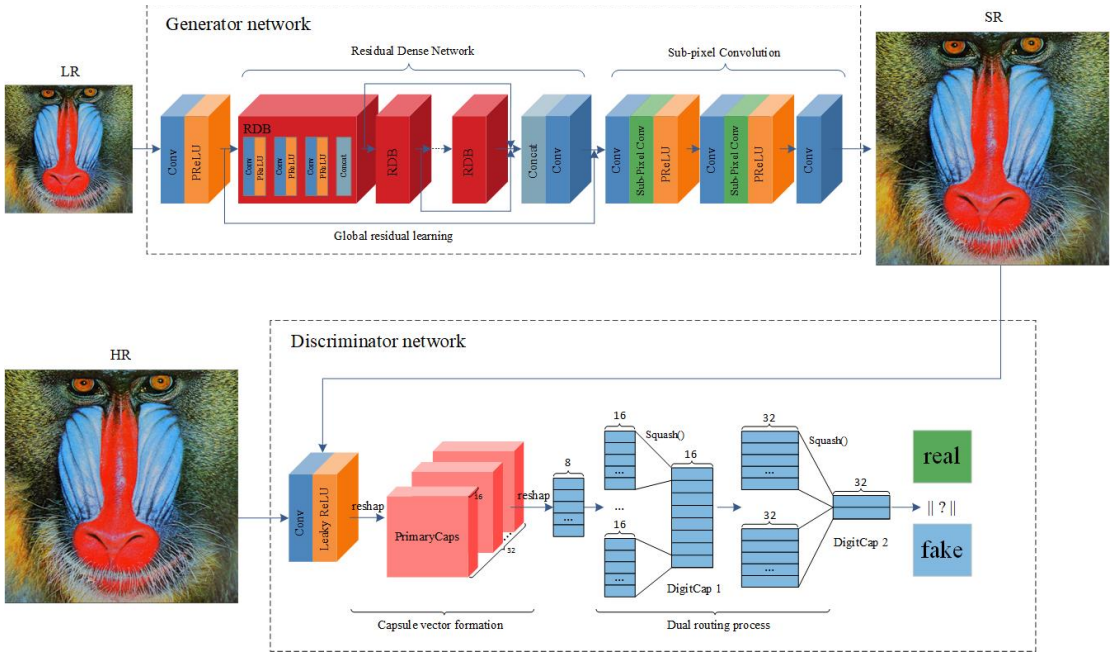


Figure 1. Structure diagram of CSRGAN.

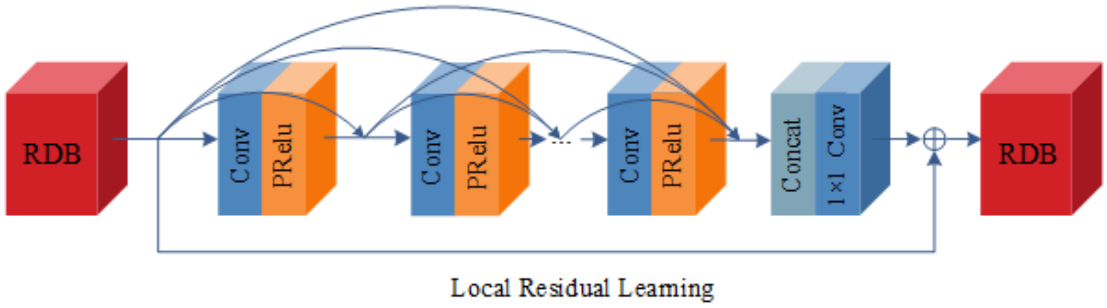


Figure 2. Structure diagram of RDB.

III. METHOD

This paper constructs a capsule generation adversarial super-resolution reconstruction network CSRGAN based on the capsule network. The network framework includes two parts: the generator network and the discriminator network. The structure of the network is shown in Fig. 1.

A. Generator network

The generator network contains two parts: a dense residual network layer and a sub-pixel convolutional layer. The dense residual network contains 16 residual blocks composed of a convolutional layer and an activation function—each residual dense block structure, As shown in Fig. 2. using this structure can avoid the disappearance of the gradient during backpropagation while using Global Residual learning to obtain global dense features from the original LR image, and further combine the shallow features and deep features to fully invoke the layering of low-resolution images features to improve feature extraction capabilities so that the network learns more effective features.

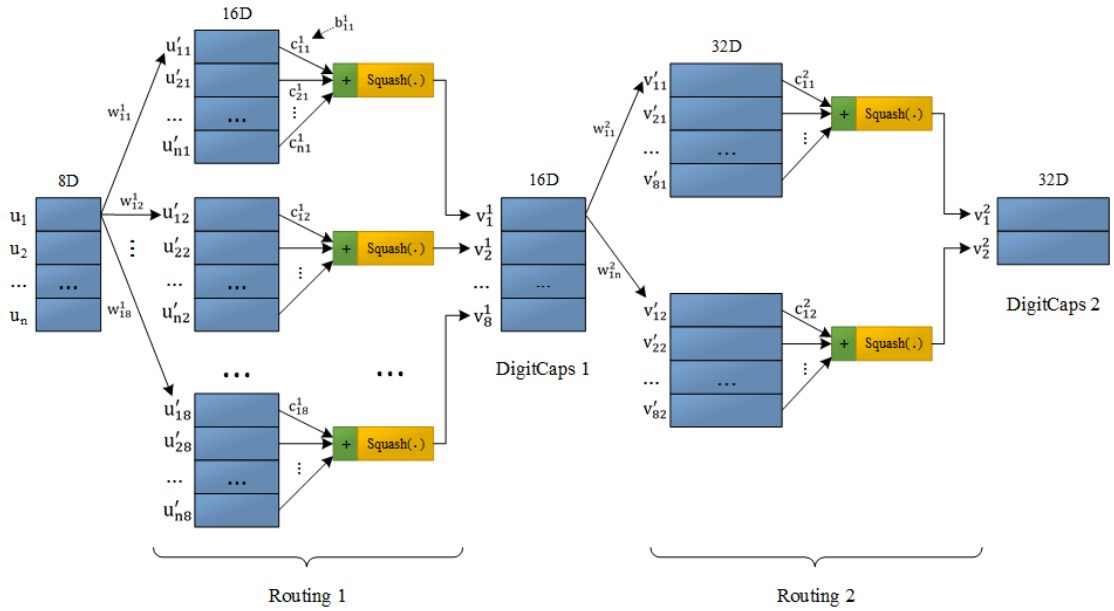


Figure 3. u is n 8-dimensional vectors obtained by encapsulating and flattening the data through the PrimaryCaps layer, v^1 is a 16-dimensional digital capsule vector after a routing process and v^2 is the 32-dimensional vector output at the end of the second routing process. w_{ij} is the fully connected weight matrix of the connected vector. The vectors u' and v' are the prediction vectors of u and v^1 obtained through a linear hierarchical relationship, the subscripts i , j correspond to the number of vector u and the next layer vector v^1 , respectively. The parameter b_{ij} is the log prior probability from the lower capsule to the upper capsule. The parameter b_{ij} is the logarithmic prior probability from the low-level capsule to the high-level capsule and is used to update each c_{ij} correspondingly. c_{ij} is the coupling coefficient connecting the two layers' vectors before and after, representing the degree of correlation between the i-th vector of the previous layer and the j-th vector of the next layer. Squash is a linear rectification function, which normalizes the results obtained.

B. Discriminator network with double routing

The discriminator network uses a dual-route capsule network to achieve true or false classifications, including a capsule vector forming part and a dual-routing part. The capsule vector forming part includes a Conv layer, a PrimaryCaps layer, and two DigitCaps layers.

The capsule vector formation part is that after the Conv layer performs preliminary feature extraction on the input image, the original scalar neuron of the convolutional network is repackaged into a vector neuron during the parameter transfer process PrimaryCaps layer. The length of each vector represents the estimated probability of whether the object exists, and its direction records the posture parameters of the object.

After the PrimaryCaps layer, the obtained multi-dimensional tensor is flattened into a one-dimensional array and then input into the two Routing processes to obtain two DigitCaps layers, respectively. These two DigitCaps layers implement the feature's coarse extraction process and the feature's fine extraction process to classify the features. There is a more gradual transition, which is beneficial to improve the classification accuracy. The dual routing is shown in Fig. 3.

It should be emphasized that the importance of low-level feature capsules to high-level feature capsules can be measured by c_{ij} and the corresponding prediction vector. c_{ij} is the normalized result of b_{ij} . The formula is as follows:

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ij})} \tag{1}$$

The update of c_{ij} is affected by b_{ij}. The initial value of b_{ij} is 0. The c_{ij} normalized by the formula (1.1) is assigned the average probability value, such as: $c_{11}^1 + c_{12}^1 + ... + c_{18}^1 = 1$.

The update of b_{ij} uses the following formula:

$$b_{ij} = b_{ij} + \hat{u}_{ij} \cdot v_j \tag{2}$$

Relying on (1) and (2) the capsule network can iteratively improve the coupling coefficient c_{ij}.

During training, the input of the capsule discriminator is the generated high-resolution image and the original high-resolution image of the training set. Its output is the binary classification result calculated by the second normal form of the vector output from the second DigitCaps layer, corresponding to the generated image (fake) and real image (real).

The dimension expansion of the feature vector through v^1 and v^2 corresponds to the rough extraction and fine extraction of the dual routing process. That allows the dual-route capsule network to deepen the network's understanding of image features in a higher dimension and enhance the discriminant network's discriminative ability to assist the generation of the network better.

IV. LOSS FUNCTION

This article is based on GAN network construction. The overall network loss function depends on the GAN loss function and the vector inner product loss function we proposed. To achieve a complete network function, we need to discriminant network, using the training's discriminator loss function and generator loss function [25].

A. D Loss

Since the discriminator network is built by the capsule network as a whole, and the final actual output result is a multi-dimensional capsule vector, the impact of the capsule network's loss function needs to be considered.

L_M is the edge loss function used by the capsule network to implement parameter training, which is defined as follows:

$$\begin{aligned} L_M = \sum_{k=1}^{K} T_k \max(0, m^+ - \| v_k \|)^2 + \\ \lambda(1 - T_k) \max(0, \| v_k \| - m^-)^2 \end{aligned} \tag{3}$$

Among them, k is the number of classifications, and T_k is a function of the classification. $T_k = 1$ when and only when the classification of k appears, and $T_k = 0$ when it does not exist.

The loss function of the GAN discriminator is as follows:

$$\begin{aligned} \max V(D,G) = E_{x \sim p_{data}(x)}[\log D(x)] + \\ E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \end{aligned} \tag{4}$$

After combining the super-resolution task with the above loss function and making improvements based on WGAN [26,27], we can define the final discriminator loss function, as in (5):

$$\begin{aligned} L_D = \max E_{I^{HR} \sim p_{train}(I^{HR})}[-L_M(D_{\theta_D}(I^{HR}), T = 1)] + \\ E_{I^{LR} \sim p_G(I^{LR})}[-L_M(D_{\theta_D}(G_{\theta_G}(I^{LR})), T = 0)] \end{aligned} \tag{5}$$

The two parts of the formula represent the discriminative loss of the two source data, respectively. Among them, D is the discriminator network. Since the capsule network is used as the discriminator in this article, the actual output result is a vector. G is the generator network, and the output is a high-resolution image. I^{HR} represents the incoming discriminant network D is the training set high-resolution (HR) image, I^{LR} represents the incoming generation network G is the pre-processed low-resolution image (LR), The LR is processed by the G network into a high-resolution image SR and then transferred to the D network. The source is the real training set whose T value is set to 1, and the source is the generated image whose T value is set to 0.

B. G Loss

When the generator is trained, the minimum value of the generated adversarial is calculated. Since the first half of the formula has nothing to do with the G network, the second half of the (1) is used during actual training, and the T value is set to 1, as follows:

$$L_{Adversarial} = \min E_{I^{LR}\sim p_G(I^{LR})}[L_M(D_{\theta_G}(G_{\theta_G}(I^{LR})), T=1)] \quad (6)$$

In addition to the generator counter loss of (4), in order to improve the perception of the texture details of the generated image, this paper proposes a vector inner product loss function based on the capsule network as follows:

$$L^{SR}_{Vecoter} = \sum_{i}^{m}(V_i(I^{HR})\cdot V_i(G_{\theta_G}(I^{LR}))-\|V_i(I^{HR})\|^2)^2 \quad (7)$$

V represents the vector before the compressed rectification function of the capsule inner product loss function, that is, a 16-dimensional vector fromthe DigitCaps layer of the capsule network. The subscript i of V represents the number of sequences to be classified, and m is the total number of classifications. There are two classifications of true and false in the experiment, and the value of i is 0 or 1.

Therefore, combining (3) and (4), the final cost function L_G of CSRGAN generator is as follows:

$$L_G = L^{SR}_{Vecoter} + 10^{-3}\cdot L_{Adversarial} \quad (8)$$

To make the final reconstructed pixels of the generated image closer to the HR, we need to train the main constraints to focus more on the pixels' accuracy while relatively weakening the dependence on the "creativity" of the adversarial network. The adjustment coefficient in (8) is set to 10-3.

V. EXPERIMENTS

A. Experimental results

The DIV2K data set is used in the experiment, which contains 800 images of the training set, 100 images of the test set, and 100 images of the verification set. Random horizontal flipping is used to enhance the data set. The bitch size of the network is set to 16. Before the experiment, it is necessary to obtain the LR image by bicubic interpolation of the HR and then use the super-resolution reconstruction algorithm to reconstruct the LR image into an SR image. The processes from LR to HR and HR to SR are all performed with a 4 times scale factor. The network's ultimate goal is to train a generator network G, which can generate a corresponding HR image from a given LR image. The experiment finally realizes the optimization of the above-mentioned mini-max problems through alternate training of D-network and G-network. The RMSProp algorithm, which is more suitable for gradient instability, is used for parameter optimization during training. In order to make the discriminator have a certain discriminative performance in the experiment, the experiment is set to start training the G network when the D network is trained 25 times. After many experiments, it is found that the training ratio of the D to G network is adjusted to 3:1. The two networks can have a better convergence effect.

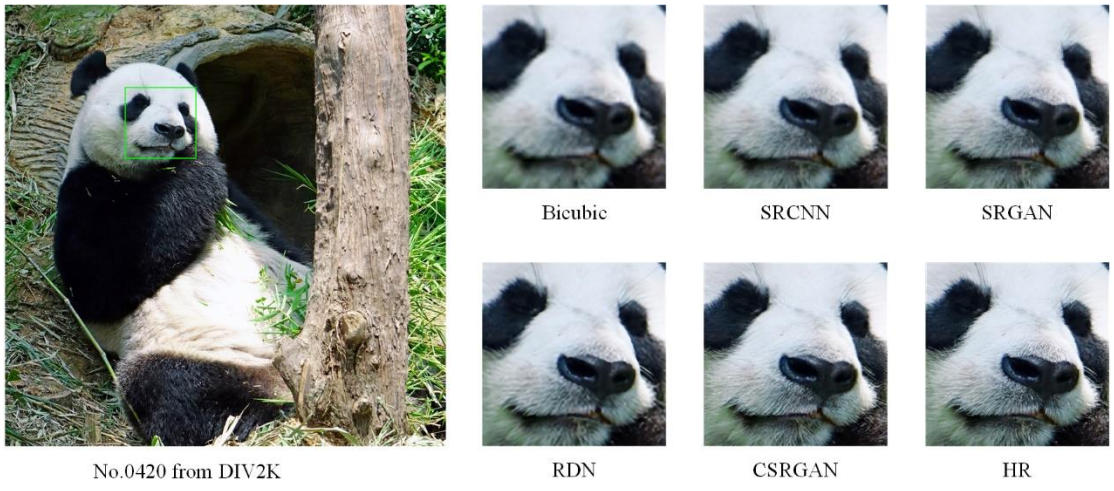


Figure 4. Panda.

Figure 5. Rome.

Figure 6. Nut.

The experimental content mainly focuses on the three dimensions of image texture effect, image brightness, and local features. The comparison with Bicubic, SRCNN, SRGAN, and RDN is shown in figures below:

B. Analysis of results

Fig. 4, Fig. 5 and Fig. 6 respectively capture some of the 250×250 pixels, 185×185 pixels, and 100×100 pixels corresponding to the experimental results.

In pixel-level reconstruction tasks, the texture is an important part of describing the image's details, usually in the reconstruction effect of the partial lines of the image. It can be seen from the experimental results that the reconstructed image obtained by the Bicubic method is blurred, especially in the image where the pixel difference is dense, and a lot of texture details are missing. For example, in Fig. 6, the lines on the surface of the reconstructed fruit in the image are even disconnected and discontinuous. The image reconstructed by SRCNN cannot well restore the image's texture, and there is a lack of smooth transition between local edge pixels in the high-multiple reconstruction task. For example, in Fig. 5, the image reconstructed by SRCNN does not accurately restore the straight lines of smaller wall tile targets, and some curves are reconstructed into straight lines.

In the experimental evaluation of super-resolution reconstruction, brightness is an important parameter of the similarity index of image structure. It is also the image element that

human visual senses are most likely to perceive. As shown in Fig. 5, SRGAN, which uses GAN as the network structure, has a significant deviation from the label data in the restoration of the brightness space. The reconstructed building image produces a generally high brightness result. On the other hand, although the sharpness of SRGAN is greatly improved compared with SRCNN, the general grid-like texture shown in Fig. 6 is generated in the image after reconstruction, which also greatly impacts the look and feel of the image. RDN has high clarity, but it will produce excessively sharp edges in the image's local details. As shown in Fig. 5, the reconstructed image's wall lines are excessively sharp, which is not consistent with the style of ancient Roman buildings, which makes the image not realistic in appearance. CSRGAN significantly weakens the raster effect produced by the reconstruction of the SRGAN architecture and has richer local texture details, ensuring that the image is clear enough during high-magnification reconstruction.

Unlike other reconstruction methods that lack attention to the correlation between local features, CSRGAN using the capsule network is more in line with the natural distribution of the image in terms of local features. As shown in Fig. 4, we can see that in the reconstruction results of the panda's nose and surrounding hair, CSRGAN has the most natural and smooth transition and can achieve more accurate reconstruction in different parts of the hair with similar color and texture. As shown in Fig. 6, in the original image of HR, the lines and grooves on the nut's surface also have a large number of different parts of the same color but different color depths. CSRGAN also accurately achieves the reconstruction and has a look and feel closer to the real image.

In the experiment, we use PSNR as the key image evaluation index. The PSNR training results of CSRGAN based on the Set5 dataset are shown in Fig. 7.

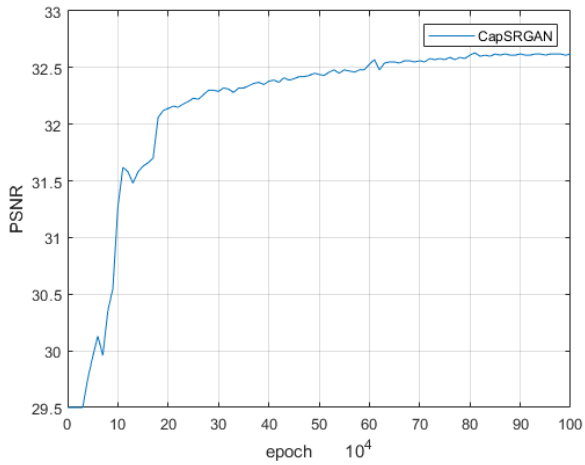


Figure 7. Train PSNR is evaluated on Set5.

As shown in Fig. 7, CSRGAN has a relatively stable convergence process and can stably generate high-quality super-resolution images after reaching a certain number of training times.

In this experiment, the two standards of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) comprehensively evaluate the results of each experimental method as follows:

TABLE I. PSNR AND SSIM

Method	Bicubic	SRCNN	SRGAN	RDN	CSRGAN
PSNR	28.39	30.45	29.43	32.44	32.58
SSIM	0.8102	0. 8616	0.8477	0.8988	0.9003

It can be seen from Table I that Bicubic's PSNR and SSIM achieved 28.39 and 0.8102, respectively, which is the lowest score for this experiment. The PSNR and SSIM parameters of SRGAN are 29.43 and 0.8477, respectively, and both indicators are weaker than SRCNN and RDN. The two parameters of CSRGAN's PSNR and SSIM's evaluation value in the experiment are compared with the leading RDN reconstruction algorithm in the field. The two parameters have increased by 0.14 and 0.0015, respectively, which shows that CSRGAN has the ability to generate more realistic super-resolution images than RDN.

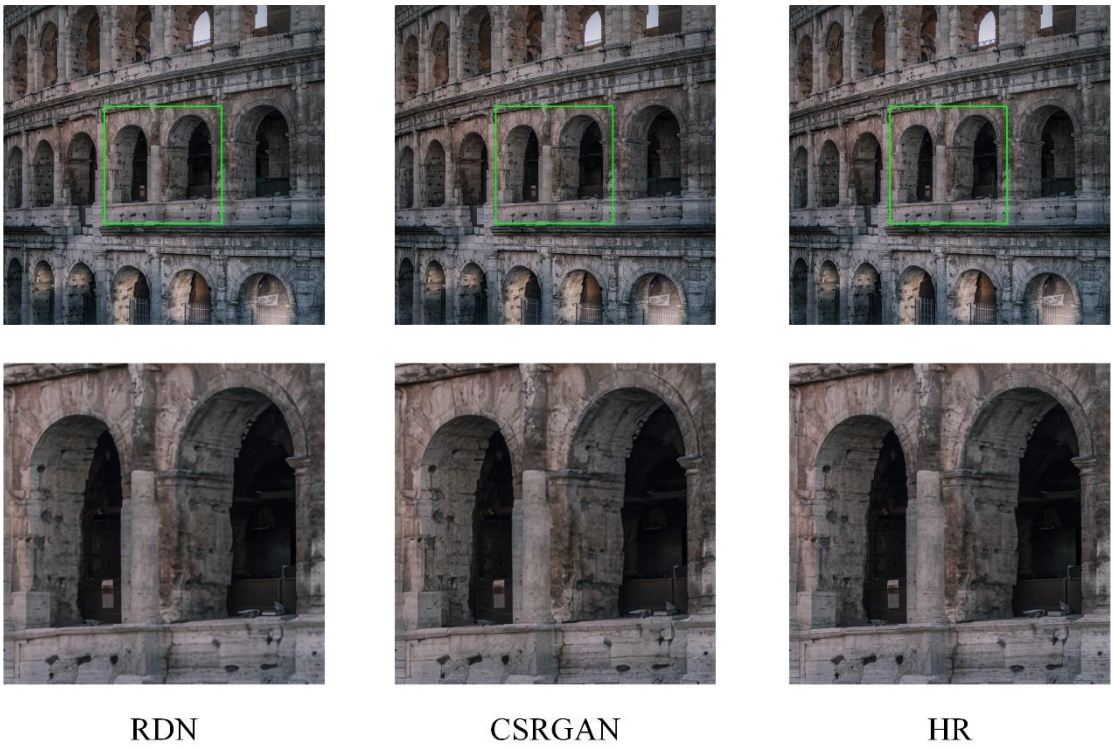


Figure 8. The Roman Colosseum(a).

Figure 9. The Roman Colosseum(b).

C. Experimental negative sample analysis

Nevertheless, CSRGAN still has a lot of work worthy of in-depth research in the future, and there are also unresolved factors that negatively affect the imaging effect. In the few images in the experiment, we also found some samples with poor performance. As shown in Fig. 8 and Fig. 9 below.

Although the images captured in Fig. 8 and Fig. 9 are from the same set of experimental data, the

image area CSRGAN shown in Example 1 exceeds the image generated by RDN in terms of the doorway's details and the overall texture.

As shown in Fig. 10 and Fig. 11, the image data is still taken from the same experimental test sample. In Fig. 10, the clarity of the reconstructed image of the trees generated by CSRGAN is far better than that of the image generated by RDN, and even the restoration of the stairs and walls in Fig. 11 gives the audience a clearer subjective feeling. However, we can observe the details of the bricks of the steps and the wall's decoration, which produces a larger deviation compared with the HR image.

We believe that the situation in Fig. 8 and Fig. 9 is highly likely to be caused by the fact that the capsule network has few relevant features in the training set and weak local correlation (such as the strong correlation between the overall features of the human face and the features of local facial features), and that the level of the capsule network in CSRGAN is relatively shallow and no deeper features are extracted. Improving the capsule network's performance by deepening the capsule network is a direction of our future related research. In addition to the above factors, some of the influencing factors in Fig. 10 and Fig. 11 may come from the GAN model. The selection of effective features between capsule layers is achieved through clustering algorithms, which will slow down the overall training speed during training. Therefore, how to optimize the capsule network clustering algorithm and improve its training speed in the future is also an important research direction in the research of super-resolution reconstruction of capsule network.

Figure 10. castle(a).

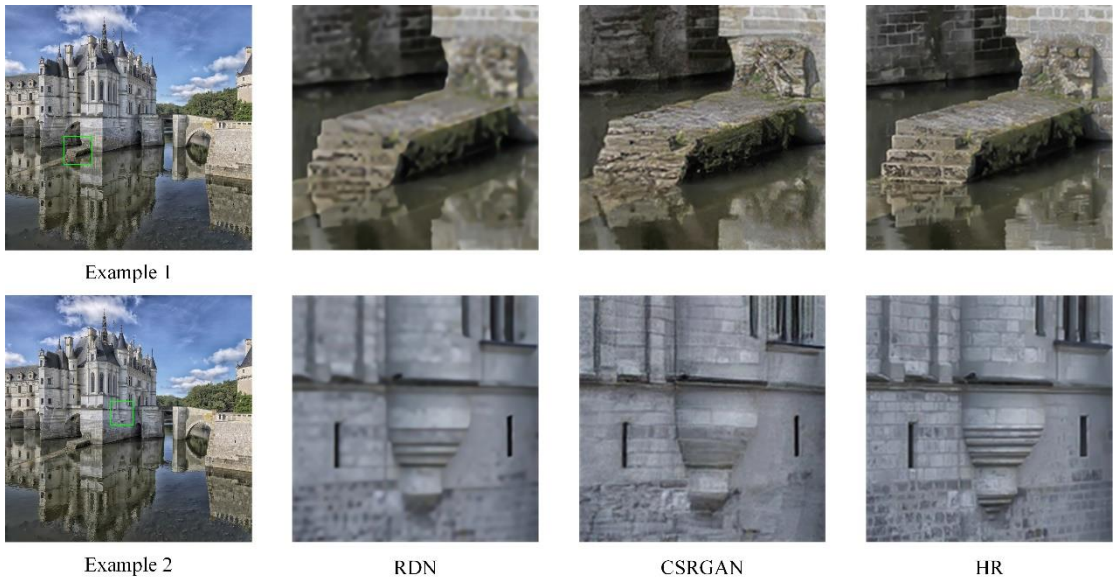


Figure 11. castle(b).

VI. CONCLUSION

To improve the image quality of super-resolution reconstruction of images and obtain better local detail texture performance, this paper proposes a GAN model CSRGAN that uses a capsule network as a discriminator. This model uses a capsule network to replace the traditional CNN discriminator network, which enables the network model to have a more detailed grasp of local spatial information, and enables the generation network to achieve the more accurate reconstruction of regions with complex local pixel distribution; Simultaneously, the vector inner product loss function is added, and the images generated by CSRGAN achieve better detail performance through the network's training.

In summary, although there are still many areas for improvement in the network, CSRGAN is still a more effective attempt. The final PSNR and SSIM results of the experiment show that the CSRGAN model has a better reconstruction effect than the CNN-based super-resolution model on the DIV2K data set, which proves that the attention of capsule network to local spatial information in images is effective, which will contribute to the generation of images with higher reduction accuracy in super-resolution reconstruction tasks.

REFERENCES

[1] Lu H, Li Y, Chen M, et al. Brain intelligence: go beyond artificial intelligence [J]. Mobile Networks and Applications, 2018, 23(2): 368-375.

[2] Lu H, Zhang M, Xu X, et al. Deep fuzzy hashing network for efficient image retrieval [J]. IEEE Transactions on Fuzzy Systems, 2020.

[3] Lu, H., Member, S., Tang, Y., & Sun, Y. (2020). DRRS-BC: Decentralized Routing registration system based on blockchain. IEEE/CAA Journal of Automatica Sinica, 1–9.

[4] Lu H, Zhang Y, Li Y, et al. User-oriented virtual mobile network resource management for vehicle communications[J]. IEEE Transactions on Intelligent Transportation Systems, 2020.

[5] Lu H, Li Y, Mu S, et al. Motor anomaly detection for unmanned aerial vehicles using reinforcement learning [J]. IEEE internet of things journal, 2017, 5(4): 2315-2322.

[6] Chen Z, Lu H, Tian S, et al. Construction of a Hierarchical Feature Enhancement Network and Its Application in Fault Recognition [J]. IEEE Transactions on Industrial Informatics, 2020.

[7] 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.

[8] H. Lu, R. Yang, Z. Deng, Y. Zhang, G. Gao, R. Lan, “Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM”, ACM Transactions on Multimedia Computing Communications and Applications, 2020.

[9] Dong Chao,Loy Chen Change,He Kaiming,Tang Xiaoou. Image Super-Resolution Using Deep Convolutional Networks. [J]. IEEE transactions on pattern analysis and machine intelligence,2016,38(2).

[10] Shi W, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1874-1883.

[11] Kim J, Kwon Lee J, Mu Lee K. Deeply-recursive convolutional network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1637-1645.

[12] Ledig C, Theis L, Huszár F, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681-4690.

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

[14] Wang X, Yu K, Wu S, et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 0-0.

[15] Kim J, Kwon Lee J, Mu Lee K. Accurate image super-resolution using very deep convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1646-1654.

[16] Wang X, Yu K, Dong C, et al. Recovering realistic texture in image super-resolution by deep spatial feature transform[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 606-615.

[17] Timofte R, Rothe R, Van Gool L. Seven ways to improve example-based single image super resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1865-1873.

[18] Zhang Y, Tian Y, Kong Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.

[19] Guo Y, Chen J, Wang J, et al. Closed-loop matters: Dual regression networks for single image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5407-5416.

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

[21] Jaiswal A, AbdAlmageed W, Wu Y, et al. Capsulegan: Generative adversarial capsule network[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 0-0.

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

[23] 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.

[24] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[25] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks [J]. arXiv preprint arXiv:1701.04862, 2017.

[26] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J]. arXiv preprint arXiv:1701.07875, 2017.

[27] Cao J, Mo L, Zhang Y, et al. Multi-marginal wasserstein gan [J]. arXiv preprint arXiv:1911.00888, 2019.

Super-resolution Image Reconstruction Based on Double Regression Network Model

Jieyi Lv
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
e-mail:ljylly150@163.com

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

***Abstract*—By learning nonlinear mapping functions from low resolution (LR) images to high resolution (HR) images, deep neural networks show good performance in image super-resolution (SR). However, the existing SR approach has two potential limitations. First, learning the mapping function from LR to HR images is usually an ill-conditioned problem, since there exist an infinite number of HR images that can be down-sampled to the same LR image. Thus, the space of possible functions can be very large, making it difficult to find a good solution. Second, paired LR-HR data may not be available in real-world applications, and the underlying degradation method is often unknown. For this more general case, existing SR models tend to generate adaptive problems and produce poor performance. To solve the above problem, we propose a dual regression scheme that reduces the space of possible functions by introducing additional constraints on LR data.**

Keywords-Super Resolution; Mapping Function; Deep Neural Network; Double Regression Model

I. INTRODUCTION

Image super-resolution reconstruction is to reconstruct high resolution image from low resolution image. This typical problem has been widely used in astronomy, physics, medicine and other fields. In the initial super-resolution reconstruction, interpolation methods are used, including nearest neighbor interpolation, bilinear interpolation, bicubic interpolation and so on. The interpolation method can effectively enhance the image resolution with a small amount of computation, but simply increasing the number of pixels will lead to blurred edges of the image. In addition, methods such as image super-resolution, local linear regression, dictionary learning and random forest based on sparse coding have also been widely used in many fields.

Single image super resolution is an ill-conditioned inverse problem: one LR image can correspond to multiple HR images, and the reconstructed HR images often have defects such as detail loss, edge aliasing and blurring. Deep learning technology has greatly promoted the rapid development of the field of computer vision. The vast majority of current SISR algorithms are based on end-to-end deep learning technology, that is, they directly learn the mapping between LR and HR. Although deep learning-based SISR method has made great progress, it still has problems in practical application: better results tend to rely on deeper networks, and more parameters also require more training data. This requires longer training and reasoning times, as

DOI: 10.2478/ijanmc-2022-0039

well as greater computing power and memory. As a result, its usefulness is greatly limited, especially in resource-constrained mobile devices.

In this paper, a new dual regression scheme is proposed to form a closed loop to improve SR performance. We introduce an additional constraint to reduce the possible space so that the super-resolved image can reconstruct the input LR image. Ideally, if the mapping from LR→HR is optimal, the super resolution image can be d down-sampled to obtain the same input LR image. With such constraints, we can estimate the underlying down-sampling kernel, thereby reducing the space of possible functions and finding a good mapping from LR to HR. The specific scheme is shown in Figure 1

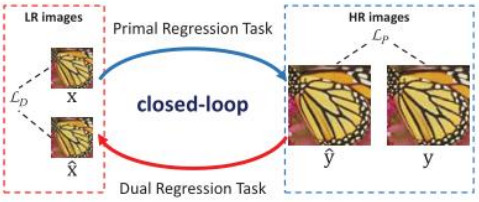


Figure 1 Composition of double regression model

II. RELATED WORK

A. Attention Mechanism

Attention mechanisms stem from studies of the human visual system. Human vision tends to focus on salient areas and ignore useless information, which increases the efficiency with which the brain processes information. The compression-expansion module proposed by Hu et al. using the attention mechanism can explicitly model the interdependence between feature channels, automatically obtain the importance of each feature channel through learning, and then suppress the less useful features according to the importance, so as to improve the performance of image classification network. The RCAN method proposed by Zhang et al. introduces the attention model of Hu et al. The channel attention module can adaptively select feature channels with richer information, which improves the performance of SISR. However, the feature channels of this attention mechanism are independent of each other, which limits the flow of feature information between channels.

B. Recursive Learning

Increasing the network depth can generally improve the performance of SISR. However, deeper networks, which generally have more parameters, require more training data. In practice, the acquisition of training data is often limited, and the risk of overfitting of network training is also increasing. One advantage of a recursive network is that it can increase the depth of the network without increasing the number of parameters. DRCN introduces recursion techniques into SISR methods for the first time: a convolutional layer is used as a recursion unit, and the weights are shared among the recursion units. DRRN improves the use of recursive techniques: residual blocks are used as recursive units, and parameters are shared among residual blocks, which also improves the performance of DRCN. Similarly, MemNet, proposed by Tai et al., also uses residual blocks as recursive units to build the network.

III. RESEARCH CONTENTS OF THE THESIS

A. Theoretical basis of dual regression model

The image super-resolution under the dual regression model focuses on the input LR image to achieve the purpose of restoring HR image. In principle, LR image I_{xLR} Can be expressed as the output of HR image after degradation function.

$$I_{xLR} = d(I_{yHR}, \partial) \tag{1}$$

Where d is the SR degradation function responsible for converting HR image into LR image, I_{yHR} Is the input HR image, which is the reference image, and the input parameters that describe the degradation function of the image. The degradation parameters are usually the scaling coefficient, blur type, and noise. In practice, the degradation process and the dependent parameters are unknown and usually only LR images are used to obtain HR images via SR methods. The SR process is responsible for predicting the inverse of the degradation function d.

$$g(I_{xLR},\delta) = d^{-1}(I_{xLR}) = I_{yE} \approx I_{yHR} \qquad (2)$$

Where g is the SR function and describes the input parameters of the function g , I_{yE} is corresponds to the input I_{xLR} of the estimated HR image. It is also worth noting that the super-resolution function in Eq. (2) is ill-posed because the function g is a non-mapping function. As a result, I_{yE} not the only one. There are many possibilities. The degradation process of the input LR image is unknown, and this process is affected by many factors such as sensor-induced noise, artifacts due to lossy compression, speckle noise, motion blur, and misfocused images. In most studies, a single down sampling function is used as the image degradation function.

$$d(I_{yHR},\partial) = (I_{yHR})\downarrow_{S_f},\{s\} \subseteq \partial \qquad (3)$$

Thereinto，$\downarrow_{s_f}$ represents the down-sampling operation, S_f Is the sampling factor. One of the down-sampling functions often used in SR is bicubic interpolation with antialiasing. In order to simulate a more realistic environment, researchers use more operations in the down-sampling function, and the overall down-sampling operation is:

$$d(I_{yHR},\partial) = (I_{yHR} \otimes \kappa)\downarrow_{S_f} + n_\sigma, \{\kappa, s, \sigma\} \subseteq \partial \qquad (4)$$

Where $I_{yHR} \otimes \kappa$ represents HR image, I_{yHR} convolution κ with fuzzy kernel, n_σ Is additive white Gaussian noise with σ standard deviation. The degradation function defined in Eq. (4) is closer to the actual function because it considers more parameters than the simple down-sampled degradation function. The purpose of image super resolution is to minimize the loss function, as follows:

$$\hat{\phi} = [\min L(I_{yE}, I_{yHR})]_\phi + h\psi(\phi) \qquad (5)$$

Where $L(I_{yE}, I_{yHR})$ is the loss function between the output HR image after SR and the actual HR image, $\psi(\phi)$ Is the regularization term. The most commonly used loss function in SR is based on the pixel mean square error, which can also be called the pixel loss. Figure 2 is a schematic diagram of the super-resolution reconstruction process.

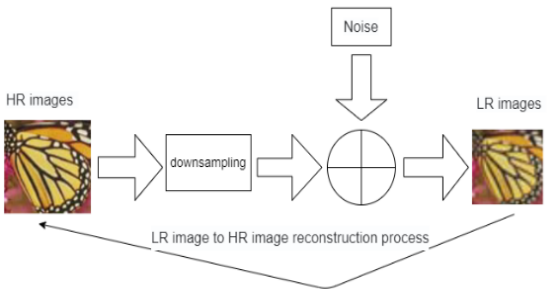


Figure 2 Super resolution reconstruction process

B. Recursive learning

One of the learning strategies based on the basic network is to recursively learn high-level features using the same module. This approach also minimizes model parameters, since the strategy requires only one module to be recursively updated, as shown in Figure 3.

One of the most commonly used recursive networks is the deep recursive convolutional network. With one convolutional layer, the DRCN can reach a receptive field of 41×41 without additional parameters. The deep recursive residual network uses the residual module Res Block as a part of the recursive module, which has a total of 25 recursions and performs better than the baseline Res Block. In addition to end-to-end recursion, the researchers also used a two-state recursive network, which shares signals between LR images and generated HR image states within the network. In general, while reducing parameters, recursive learning networks can learn complex representations of data at the cost of computational performance. In addition, increased computational requirements may cause gradients to explode or disappear. Therefore, recursive learning is often used in conjunction with multi-supervised or residual learning to minimize the risk of gradient explosion or disappearance.

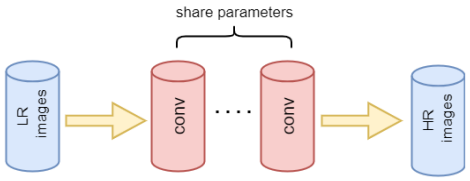


Figure 3 Recursive learning

C. Dual regression scheme for paired data

Existing methods only focus on learning the mapping from LR to HR images. However, the space of possible mapping functions can be very large, making training very difficult. To solve this problem, we propose a dual regression scheme by introducing an additional constraint on LR data. Specifically, in addition to learning the LR→HR mapping, we also learn the inverse/dual mapping from the super resolution image to the LR image. We learn to reconstruct HR images with the original map P and LR images with the dual map D simultaneously. Note that the dual map can be viewed as an estimate of the underlying down-sampled kernel. Formally, we formulate the SR problem as a binary regression scheme involving two regression tasks. The specific expression form is shown in Figure 1.

The original regression task: We need to find a function $P: X \to Y$ Make the prediction P (x) similar to its corresponding HR image Y.

Dual regression task: We seek a function $D: Y \to X$, so that the prediction of D(y) is similar to the original input LR image x.

Primal and dual learning tasks can form a closed loop to provide information supervision for training models P and D. If P (x) is the correct HR image, then the down-sampled image D(P (x)) should be very close to the input LR image x. With this constraint, we can reduce the space of functions of possible maps, making it easier to learn better maps to reconstruct HR images. By jointly learning these two learning tasks, we propose to train the super-resolution model as follows. Given N pairs of samples $S_p = \{(x_i, y_i)\}_{i=1}^{N}$

Where x_i and y_i denotes those in the paired data set i for low resolution and high resolution images. The training loss can be written as

$$\sum_{i=1}^{N} \zeta_P(P(x_i), y_i) + \lambda \zeta_D(D(P(x_i)), x_i) \quad (6)$$

Where ζ_P and ζ_D represent the loss functions of primal and dual regression tasks, respectively. Here, λ controls the weight of the dual regression loss.

D. Dual regression scheme for unpaired data

We consider a more general SR case where there is no corresponding HR data with real-world LR data. More critically, the degradation methods of LR images are often unknown, which makes this problem very challenging. In this case, existing SR models often produce serious adaptation problems. To solve this problem, we propose an efficient algorithm to adapt the SR model to the new LR data. Dual regression maps learn the underlying degradation methods and do not necessarily depend on HR images. Therefore, we can directly learn from unpaired real-world LR data for model adaptation. In order to ensure the performance of HR image reconstruction, we also add the information of pairwise synthetic data which is very easy to obtain. Given M unpaired LR samples and N paired synthetic samples, the objective function can be expressed as follows.

$$\sum_{i=1}^{M+N} \iota_{S_p}(x_i)\zeta_D(P(x_i)), y_i) + \lambda\zeta_D(D(P(x_i)), x_i) \quad (7)$$

Where, $\iota_{S_p}(x_i)$ is an index function, if $x_i \in S_p$, the function is equal to 1, otherwise the function is equal to 0.

IV. EXPERIMENT AND RESULTS

The hardware environment of this section is as follows: CPU frequency is 3.6GHz, memory is 32GB, graphics card is NVIDIA RTX 2080Ti; The software environment is: Windows operating system, Pytorch deep learning framework, Python3.7 programming language. In the experiment, 800 high-quality training images from the public dataset DIV2K were used as the training set. In the test phase, Set5, Set14 and BSD100 were used as the standard test set, which all contained various animals and plants as well as natural scene images.

Before training, the training set is enhanced by horizontal flip and Angle rotation. Considering that putting a single image directly into the model would increase the amount of computation, which is not convenient for training, each image was randomly cropped to obtain 256×256 image blocks. LR images in the training set and test set were obtained by bicuxic interpolation in Matlab language. In order to evaluate the model performance, the PEak-to-noise ratio (PSNR) and structural similarity are calculated on the Y channel (i.e., brightness) of the YCbCr channel in this experiment. The algorithm in this paper has two models in total. The generative network model Ours P for PSNR value generates the final model Ours for adversarial network training. The experimental results of the open datasets Set5, Set14 and BSD100 are respectively compared with those of the classical algorithms FSRCNN, SRGAN and ESRGAN. In the experiment, the number of modules in the intermediate feature layer of SRGAN and ESRGAN is set to 24, which is consistent with the original paper. The number of ECB modules in the proposed algorithm is set to 8, and the number of input channels is 24. The PSNR and SSIM values and the number of parameters at ×2 and ×4 magnifications were compared respectively.

Considering that the training of generative adversarial network should go through two training processes, in order to save time, only the generative network model Ours oriented to PSNR value is trained. It can be seen that the deeper the

network, the higher the objective evaluation index, but at the same time, the longer the training time. At the same time, it can be seen from the table that the model continues to increase the number of network layers in a deeper situation, which does not significantly improve the objective evaluation index, but increases the number of model parameters. Therefore, the algorithm in this paper is finally set as 8.

TABLE I. COMPARISON OF THE RESULTS OF PSNR RECONSTRUCTION BY TWO TIMES OF EACH ALGORITHM

Datasets	FSRCNN	SRGAN	ESRGAN	Ours
Set5	32.40	33.19	33.46	34.68
Set14	29.52	29.56	29.47	30.56
Bsd100	26.74	29.47	29.78	29.78

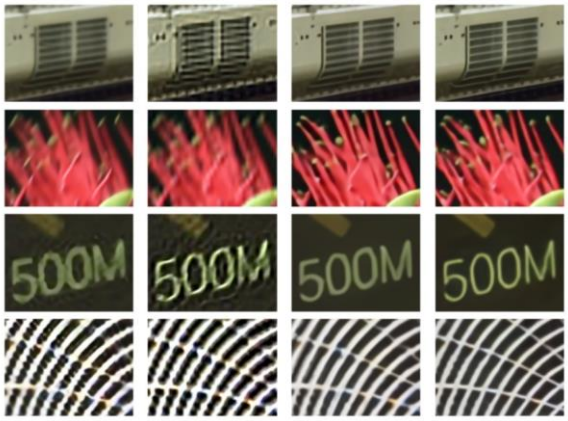
TABLE II. COMPARISON OF SSIM RESULTS OF FOUR TIMES RECONSTRUCTION OF EACH ALGORITHM

Datasets	FSRCNN	SRGAN	ESRGAN	Ours
Set5	0.8657	0.8657	0.8569	0.6895
Set14	0.7564	0.7369	0.8697	0.7698
Bsd100	0.7156	0.6689	0.6783	0.6856

Figure 4 and 5 shows the super-resolution image reconstructed by the proposed method and other advanced methods at ×4 magnification. The reconstruction results obtained by the proposed method are significantly better than those obtained by other methods. The first group of pictures recovered by FSRCNN method are blurred and have serious distortion; the second group of pictures recovered by SRGAN method have serious double shadow phenomenon around the branches; and the first group of pictures recovered by ESRGAN method also have serious blur phenomenon. The proposed method almost perfectly recovers the images under ×4 magnification. The recovery results of this method are similar. Compared with the current state-of-the-art, more lightweight network, the proposed method can better recover the image contour in the Set14 dataset at ×4 magnification. It is easy to observe that the reconstruction results in this paper are also significantly better than the other two cutting-edge methods.

(a) FSRCNN (b) SRGAN (c) ESRGAN (d) Ours

Figure 4 Double magnification reconstruction comparison of each algorithm



(a) FSRCNN (b) SRGAN (c) ESRGAN (d)Ours

Figure 5 Comparison of four times magnification reconstruction of each algorithm

V. SUMMARY

This paper presents a new binary regression scheme for paired and unpaired data. On pairwise data, we introduce an additional constraint by reconstructing LR images to reduce the space of possible functions. As a result, we can significantly improve the performance of SR models. In addition, we focus on unpaired data and apply the dual regression scheme to real-world data. We performed ablation studies with the dual regression protocol, and the model with the dual regression protocol performed better on all datasets compared to baseline. These results show that the dual regression scheme can improve the reconstruction of HR images by introducing additional constraints to reduce the space of mapping functions. We also evaluated the influence of our dual regression scheme on other models. Compared with other classical algorithms, we compared PSNR results, SSIM results and intuitive dataset images from two levels of double and four times magnification. It can be seen that the improved algorithm is significantly better than the classical algorithm. Due to the large number of super-resolution scenes, we only selected the data sets that were biased towards buildings, trees, numbers and other relevant directions. There may be some omission problems in the algorithm adaptation of medical and other professional fields, which still needs to be improved in the subsequent research.

REFERENCES

[1] Lee S, Kim JH, Heo J P. Super-Resolution of License Plate Image via Character-Based Perceptual Loss[C]//2020 IEEE International Conference on Big Data and Smart Computing. IEEE, 2020:560-563.

[2] Yong Guo, Yin Zheng, Mingkui Tan, Qi Chen, Jian Chen, Peilin Zhao, and Junzhou Huang. Nat: Neural architec-ture transformer for accurate and compact architectures. InAdvances in Neural Information Processing Systems, pages735–747, 2019, 1.

[3] Jiezhang Cao, Yong Guo, Qingyao Wu, Chunhua Shen, Jun-zhou Huang, and Mingkui Tan. Adversarial learning with local coordinate coding. In International Conference on Ma-chine Learning, 2018, 2.

[4] Tai Y, Yang J, Liu X, et al. Memnet: A Persistent Memory Network for Image Restoration[C] //Proceedings of the IEEE international conference on computer vision. 2017: 4539-4547.

[5] Zhang Y, Tian Y, Kong Y, et al. Residual Dense Network for Image Super-Resolution[C]//Proc eedings of the IEEE conference on computer vision and pattern recognition. 2018: 2472-2481.

[6] Tao Dai, Jianrui Cai, Y ongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In IEEE Conference on Computer Vision and Pattern Recognition, 2019,6.

[7] Zhen Li, Jinglei Y ang, Zheng Liu, Xiaomin Y ang, Gwang-gil Jeon, and Wei Wu. Feedback network for image super-resolution. In IEEE Conference on Computer Vision and Pat-tern Recognition, pages 3867–3876, 2019,2.

[8] Lin Wang, Wonjune Cho, and Kuk-Jin Y oon. Deceiv-ing image-to-image translation networks for autonomous driving with adversarial perturbations. arXiv preprintar Xiv: 2001.01506, 2020,2.

[9] Li Zhang, Alok Deshpande, and Xin Chen. Denoising vs.deblurring: Hdr imaging techniques using moving cameras. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 522–529. IEEE, 2010.

[10] Timo Stoffregen and Lindsay Kleeman. Event cameras, con-trast maximization and reward functions: an analysis. In Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12300–12308, 2019, 2.

[11] Christian Reinbacher, Gottfried Munda, and Thomas Pock.Real-time panoramic tracking for event cameras. In 2017IEEE International Conference on Computational Photog-raphy (ICCP), pages 1–9. IEEE, 2017, 1, 2.

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shecht-man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion, pages 586–595, 2018, 5.

Research on Gaze Estimation Method Combined with Head Motion Changes

Xiangyi Zhan
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: zxyy129@163.com

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

***Abstract*—The line of sight reflects the focus of human attention. Gaze estimation technology has a wide range of application prospects in human-computer interaction, human emotion analysis, commercial advertising, and so on. Gaze estimation needs to be jointly determined by eye movement and head movement because the human gaze is often with the head movement. In this paper, a gaze estimation system with head movement is implemented by a monocular camera, using eye movement features and head movement posture changes to estimate the gaze point. A camera is used as the information acquisition device to extract the eye movement feature information, and the offset of head movement is estimated at the same time. The final fixation point position coordinates are obtained by compensating the fixation point position in the case of head movement, and then the fixation point position coordinates are estimated. For the compensated gaze drop point, the average pixel error on the X-axis is estimated to be 118 pixels, and the average pixel error on the Y-axis is estimated to be 136 pixels.**

Keywords-Gaze Estimation; Head Movement; Pupil Center Detection; Head Eye Movement

I. INTRODUCTION

With the continuous improvement of the level of science and technology, the way of human-computer interaction (HCI) is constantly changing. The eyes are the most important organ for the human body to obtain external information, taking the gaze position from the eyes as a way of human-computer interaction, which has the characteristics of more direct, natural, and human nature. Eye tracking technology captures the eye position information and posture of the camera to obtain the gaze direction and the position of the observation point, track the eye line of the human eye, and make the human-machine interaction through eye movement naturally and conveniently. Therefore, studying gaze estimation under head motion is of great significance.

A. Gaze Estimation

The current gaze estimation methods are mainly divided into two categories: feature-based gaze estimation and epigenetic gaze estimation. Feature-based sight estimation methods can be divided into model-based and regression-based sight estimation methods according to different implementation schemes of eye direction and eye feature mapping relationship [1].

The model-based gaze estimation method first takes the straight line obtained by connecting the pupil or iris center as the optical axis. At the same time, a fixed deviation Angle between the optical axis and the visual axis is obtained based on prior knowledge to compensate for the actual line of sight to improve the accuracy of the gaze estimation. Chen proposed a 3D gaze estimation algorithm based on single-camera face tracking [2]. The algorithm was implemented without the help of an external light source. The solving equations of the 3D eye center, pupil center, and visual axis were derived, and the fixation point was solved by one-time calibration. Regression-based gaze estimation methods usually includes the following two processes: selecting and extracting eye motion and constructing regression function to obtain the mapping relationship between the line of sight and eye movement features. Eye movement is usually obtained by connecting a fixed point unrelated to eye movement with a moving point strongly related to eye movement. The most commonly used eye movement is calculated by the corner

point and the iris center point. Yamazoe proposed a real-time gaze estimation method using specific face feature association. This method calculates the geometric relationship between the eye center and the eye radius in advance and simplifies the sight estimation method based on a 3D sight model by tracking specific facial features [3]. Valenti used the method of Isocenters for iris center location and canthus location under a single webcam and achieved a better detection accuracy in low-resolution images [4].

Appearance-based gaze estimation method extracts features from the whole image and then obtains the mapping relationship model between features and line of sight direction through training. In the early stage, scholars used manual features such as HOG or LBP to extract features and then used K-nearest neighbor [5], random regression forest [6] and support vector machine (SVM) [7], and other models to achieve line-of-sight estimation. Subsequently, with the development of neural network technology, neural network models based on deep learning have been widely applied to gaze estimation. To improve the generalization performance of appearance-based models, Wang proposed a method combining adversarial learning with Bayesian inference for the overfitting of appearance, head pose, and point estimation [8]. To solve the interference of face images with free head movement on the line of sight model mapping, Zhou proposed to input different areas of the face into the 3D line of sight estimator with adaptive weighting during head movement, which greatly improved the efficiency and accuracy of the regression model [9].

B. Head Pose Estimation

Head pose estimation is intrinsically related to visual gaze estimation, namely, the head pose can represent the direction and focus of the human eyes. Physiological studies have shown that a person's gaze prediction comes from a combination of head posture and eye orientation. There are mainly the following methods for obtaining head pose: the method based on standard template matching, the method based on the geometric relationship of key points of the face, the method based on feature regression, and the method based on manifold learning [10].

The method based on standard template matching compares the header image with the model with pose labels in the feature space [11] and calculates the similarity between the input features and the standard template set [12].

The method based on the geometric relationship of the key points of the face determines the corresponding pose according to the position of the eyes, mouth, and nose tip, which can recover the global pose change of the head from the video. Fridman used the random forest machine learning method to classify the pose of the driver's facial 56-point position relationship features [13]. The geometric method is simple, fast, and has low time complexity. However, this method has high requirements for the detection and location of facial feature points.

The multi-layer perceptual neural network is a common regression tool in the method based on feature regression. This method corresponds the output layer of the network to the discrete head posture, and the head posture estimation results are obtained by training the neural network directly. Murphy train regressors by support vector regression (SVR) [14]. Huang introduced the random forest algorithm into head pose estimation [15]. The method based on regression has good real-time performance and high accuracy, but it requires a large amount of computation and is greatly affected by the head detection and positioning results.

Based on manifold learning, Huang improved the accuracy of head pose in non-uniform samples by supervising local subspace learning; and discussed the phenomenon of over fitting [16]. Foytik roughly estimated the head pose through the linear supervision function, and then realized the accurate judgment of the head pose through the linear regression function [17]. The method based on manifold embedding has high time complexity and low accuracy, which is still far from the real practical process.

II. METHODS

The experimental environment of this study is simple, and image data of subjects are collected through a monocular camera and a single screen. The computer display is 31.5 inches with a

resolution of 2560×1440 pixels, and the camera is fixed in the middle of the top of the screen with a camera of 2048 × 2048 pixels. A total of 12 laboratory members, including 8 men and 4 women, participated in the data collection. Aged between 23 and 30 years with good health and good vision. Before the experiment, each subject was familiar with the specific contents and precautions of the experiment, and they were willing to participate in the experiment. In order to eliminate external interference, each experiment was completed by only one subject alone after making experimental preparations. The experiment was carried out in natural light.

The data collection in this study was divided into two parts. The first part was a static experiment, in which subjects were required to move their heads as close to the starting position as possible. The second part is the dynamic experiment, in which subjects' heads can move freely.

A. Static Experiment

Before starting the experiment, the camera and other equipment should be calibrated to ensure that they are in normal operation. Subjects e adjust their seated position on the chair to be approximately 60cm away from the center screen, ensuring that they are within the best focal length of the camera and that the origin of the subject's head is as close to the origin of the camera coordinate system as possible.

At the beginning of the experiment, a red dot with a radius of 25 pixels was displayed on the screen. The red dot represented the object to be captured by the eye. Each time the red circle appeared, the subject was asked to look at the center of the red circle and press the space bar at the same time. At this point, the camera would take an image of the subject's face, and the software would record the coordinates of the target point, which would then show the next origin. This is shown in Figure 1. Through the experiment, a set of data, including a face image and screen coordinates, are obtained.

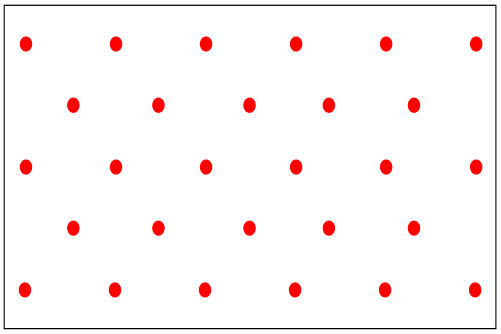


Figure 1. Schematic diagram of gaze target in static experiment

When the head is still, there is only eye movement in the process of gazing at the target, and the images captured are all frontal face images, which are almost the same, except for the difference in eye direction, as shown in Figure 2.

Figure 2. Partial data of static experiment

B. Dynamic Experiment

Experiments in the dynamic line-of-sight estimation stage are called dynamic experiments, and the experimental environment is the same as that of static experiments. Before starting, subjects should adjust their sitting position so that the distance between subjects and the center screen is about 60cm. At the beginning of the experiment, the subjects' eyes were required to face the center of the middle screen, and the initial state of head posture was calibrated by pressing the space bar. Dynamic experiments require subjects to move their heads in a specific way by focusing their eyes on a static point in the center of a computer screen, and moving their heads while still focusing on that point.

In this experiment, the subject can rotate the head freely, rotate around the X, Y, and Z axes, and generate the pitch Angle, yaw Angle, and roll Angle as well as the translation along the X, Y, and Z axes. A 5-second video clip was collected for this experiment. During head movement, the coordinated movement of the eye and the head occurs during the fixation of the target, and the pictures obtained by shooting show various appearances, resulting in the attitude change of the eye region, as shown in Figure 3.

Figure 3. Partial data of dynamic experiment

III. METHODS

In this study, we used a feature-based gaze estimation method to estimate the subject's fixation position by shooting video images with a non-wearable monocular camera. For the estimation of head motion and pose, we establish a rigid body model of the head and estimate it by geometric calculation. In this study, we use image processing and feature extraction methods to fuse head features and eye features to realize the research of gaze estimation. The technical route for the construction phase of this article is shown in Figure 4.

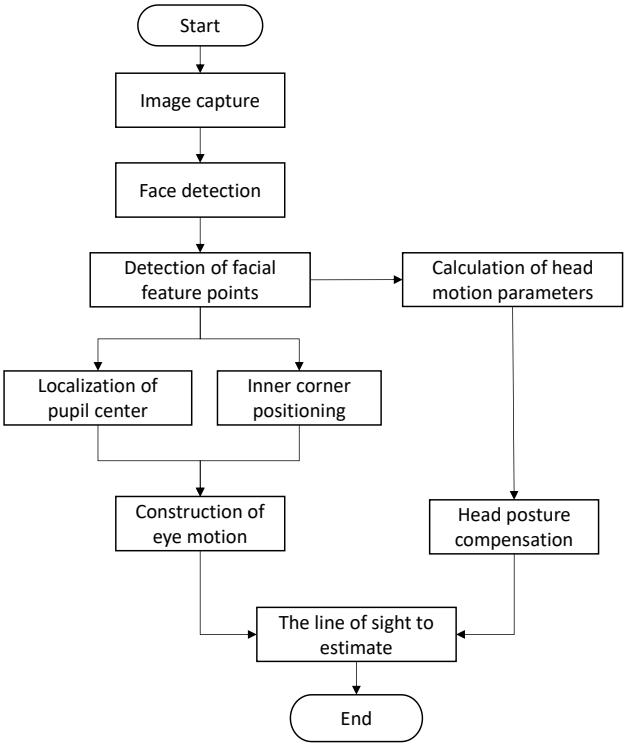


Figure 4. Technology route

A. Eye Feature Extraction

To obtain the eye image, the location of the iris center of the human eye and the inner corner of the human eye are extracted to construct eye motivities. We need to extract the position region of the face and the position coordinates of the key points of the face from the image.

The face detection method based on Adaboost and wavelet features are used to obtain the location region information of the face. Based on Haar-like features, the algorithm added a cascade classifier and integral graph algorithm to improve the speed of image processing, meet the real-time operation requirements of the whole fixation point estimation algorithm, and also to ensure the accuracy of face detection.The theory of integral graph is introduced into the algorithm, which can reduce the problem of excessive computation due to the excessive number of Haar features in the search window. The result of face recognition is shown in Figure 5.

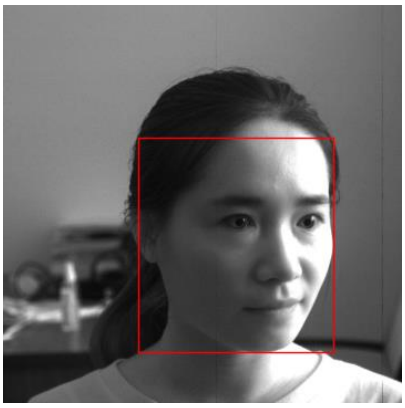


Figure 5. Face detection

Next, face feature point detection, face feature point detection is also known as face alignment, through the algorithm to detect the inner and outer corner of the face, nose tip, corner of the mouth, and another key points of the face position coordinates. Under normal circumstances, we get the position coordinates of 68 feature points of the face. After using the adaboost+haar-like algorithm to obtain the region location of the face, this paper uses the dlib c++ library to detect the feature points of the face. The position coordinates of 68 feature points of the face were obtained. The main feature points of the face included the feature points around the eye area, the coordinate points of the inner corner of the eye needed to constitute the eye movement, and the important feature points of the face needed to calculate the rotation matrix and displacement matrix of the head. The detection results of face feature points are shown in Figure 6.

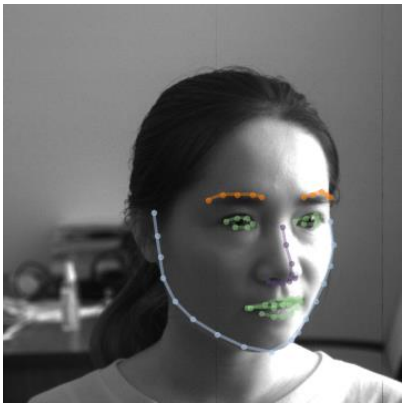


Figure 6. Facial feature points

Based on face feature points, this paper uses gradient analysis combined with star rays to calculate the position coordinates of iris center points. The algorithm can be run in the case of low-quality images, which not only ensures the computational speed but also improves the robustness of the algorithm, and can meet the different head movements and the changes of light and shade.

After obtaining the eye movement information, the mapping relationship between the feature information data of eye movement and the position coordinates of points on the screen should be established. The mapping relationship between the feature data of eye movement data and the coordinate points of the screen is a kind of nonlinear mapping relationship. Due to the real-time requirement of the gaze point estimation system, it is necessary to select a mapping model that can process the high-dimensional data and has a fast training speed. After a comprehensive comparison of various mapping models, random forest is chosen as the mapping model.

B. Head Movement Posture

Through the motion analysis of the head sequence image, the global motion of the whole head is considered as a rigid body motion model, and the head motion parameters are estimated. An object whose 3-D distance between any pair of points on the object does not change with time is defined as a rigid body. The human head can be simplified into a rigid body in the physical sense, that is, the distance between the surface points of the head's rigid body remains unchanged during its movement. Therefore, motion parameters can be estimated based on the property that the distance between points remains unchanged before and after rigid body motion.

Compared with the classical head pose estimation methods, EPNP and POSIT algorithm, it is found that the head pose estimation algorithm based on PNP has smaller advantages in estimation accuracy than the POSIT algorithm, while the head poses estimation algorithm based on POSIT has obvious advantages in calculation speed. Therefore, In this paper, the POSIT algorithm is chosen to solve the head posture, which not only ensures the requirements of solving accuracy but also meets the requirements of computing speed.

The position coordinates of the obtained facial key points and the corresponding key points of the general 3D face model were mapped and projected, and the rotation and displacement matrix (RT) of the final head pose estimation and the Yaw, Pitch, and Roll angles of the final head were calculated.

68 face feature points were obtained, from which several key points with different depths of high reliability and robustness were selected to ensure the accuracy and robustness of head pose estimation. Therefore, 8 key points were selected. That is, the left corner of the left eye, the right corner of the left eye, the right corner of the right eye, the left corner of the right eye, the nose tip, the left corner of the mouth point, the right corner of the mouth point and the lower frontal corner are used to estimate the head pose. The estimation result of the head pose is shown in Figure 7.

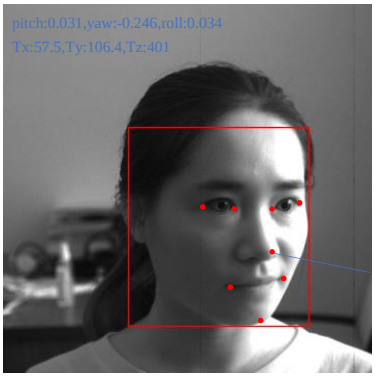


Figure 7. Estimation of head pose

C. Gaze Estimation

Is by the human eye eye eye gaze position information and joint decision movement of the head, in the head and the screen under the condition of relatively static, estimates the gaze of the initial state, to achieve the head moving cases of the fixation point estimate, need to deal with the movement of the head, above calculated the rotation matrix and translation matrix of the head, The initial fixation position was compensated according to the deflection of the tester's head, and the actual fixation position was finally obtained.

We assume that the influence of head movement on the position coordinates of fixation point is $(\Delta u_x, \Delta u_y)$.In the initial case, the head remains stationary between the screen, and the 3D initial coordinate of a point of the head is set as (x_0, y_0, z_0), the pixel coordinates of the projection of this point on the two-dimensional plane is (u_0, v_0), when the head movement occurs, the coordinates of this point are (x_1, y_1, z_1), We obtained the rotation matrix R and translation matrix T in the case of head movement through the improved POSIT algorithm, and the formula can be obtained as follows:

$$\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} = R \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + T \tag{1}$$

Where the R rotation matrix and T translation matrix are shown below.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{2}$$

$$T = \begin{bmatrix} \Delta X \\ \Delta Y \\ \Delta Z \end{bmatrix} \tag{3}$$

Therefore, the offset caused by the head movement can be calculated by the formula $(\Delta u_x, \Delta u_y)$, and f is the focal length of the camera.

$$\Delta u_x = f\frac{x_1}{z_1} - u_0 \tag{4}$$

$$\Delta u_y = f\frac{y_1}{z_1} - v_0 \tag{5}$$

The position coordinate of the gaze point under the stationary state of the head is (u_x, u_y), and the position coordinate of the gaze changed by the head movement is (s_x, s_y):

$$s_x = u_x + \Delta u_x \tag{6}$$

$$s_y = u_y + \Delta u_y \tag{7}$$

IV. Results and Analysis

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready

to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

Finally, the eye line estimation system of head movement was tested, and 12 subjects who participated in the experimental data collection participated in the test experiment. Each test participant needs to independently complete data collection, calibration training, and result testing. During the test, the test subjects were required to gaze at the 12 test points displayed on the screen in turn. To avoid interference caused by multiple test points at the same time, only one fixation point was displayed on the screen at the same time, and the calculated position coordinates of the fixation points were represented by red crosses. In the case of head movement, the data obtained from the test are sorted out, and the form of visual display of the experimental results is presented, as shown in Figure 8.

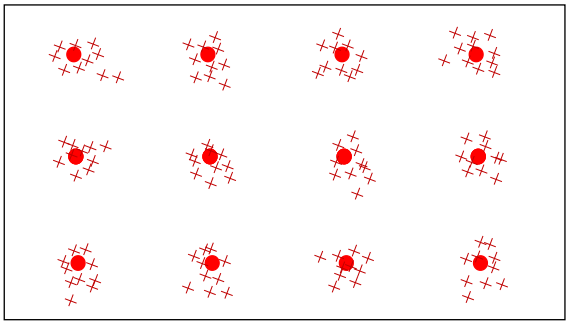


Figure 8. Schematic diagram of the test effect

To investigate the influence of the compensation effect of head motion and posture on the experimental results, the same group of testers was asked to carry out the test experiment with the same test equipment and test points under the same lighting conditions when the head was still and the head was moving. The test results are shown in the following table:

TABLE I. COMPARISON TABLE OF EXPERIMENTAL RESULTS UNDER DIFFERENT CONDITIONS

State of the head	average pixel error		average Angle error	
	X-axis (pixel)	*Y-axis (pixel)*	*X-axis (degree)*	*Y-axis (degree)*
The head rest	83	125	1.74	2.32
The head movement (No compensation)	265	380	7.23	8.15
The head movement (A compensation)	118	136	2.88	3.23

V. CONCLUSION

Gaze estimation is of great significance in the field of human-computer interaction, which is widely used in psychology, graphics, medicine, advertising psychology, military science, and other fields. At present, model-based line-of-sight estimation methods are complicated, and appearance-based line-of-sight estimation methods need a large amount of training data. In this paper, a simple gaze estimation method is adopted. Firstly, the face image is taken as the input, and the gaze estimation is performed in the initial state by extracting the eye movement information. Then, by extracting face feature points, the head rigid body motion model was established, and the head motion posture was calculated to estimate the head posture. Finally, the head movement posture was used as compensation for the eye movement gaze estimation, and the final free head movement gaze estimation was obtained.The experimental results show that the rigid motion model of the head can be used as compensation for the gaze estimation, and the head can be accurately estimated under the condition of free motion. For the compensated gaze drop point, the average pixel error of the estimated axis is 118 pixels, and the average pixel error of the Y-axis is 136 pixels.

ACKNOWLEDGMENT

This work is supported by The National Natural Science Foundation of China (No. 52072293).

REFERENCES

[1] Wood E, Bulling A. Eyetab: Model-based gaze estimation on unmodified tablet computers[C]//Proceedings of the Symposium on Eye Tracking Research and Applications. 2014: 207-210.

[2] Chen J, Ji Q. 3D gaze estimation with a single camera without IR illumination[C]//2008 19th International Conference on Pattern Recognition. IEEE, 2008: 1-4.

[3] Yamazoe H, Utsumi A, Yonezawa T, et al. Remote gaze estimation with a single camera based on facial-feature tracking without special calibration actions[C]//Proceedings of the 2008 symposium on Eye tracking research & applications. 2008: 245-250.

[4] Valenti R, Staiano J, Sebe N, et al. Webcam-based visual gaze estimation[C]//International Conference on Image Analysis and Processing. Springer, Berlin, Heidelberg, 2009: 662-671.

[5] Zhang Y, Bulling A, Gellersen H. Discrimination of gaze directions using low-level eye image features[C]//Proceedings of the 1st international

workshop on pervasive eye tracking & mobile eye-based interaction. 2011: 9-14.

[6] Wang Y, Shen T, Yuan G, et al. Appearance-based gaze estimation using deep features and random forest regression[J]. Knowledge-Based Systems, 2016, 110: 293-301.

[7] Huang Q, Veeraraghavan A, Sabharwal A. TabletGaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets[J]. Machine Vision and Applications, 2017, 28(5): 445-461.

[8] Wang K, Zhao R, Su H, et al. Generalizing eye tracking with bayesian adversarial learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 11907-11916.

[9] Zhou X, Jiang J, Liu Q, et al. Learning a 3D gaze estimator with adaptive weighted strategy[J]. IEEE Access, 2020, 8: 82142-82152.

[10] Murphy-Chutorian E , Trivedi M M . Head Pose Estimation in Computer Vision: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2009, 31(4):607-626.

[11] HE S, LIANG A, LIN L, et al. A Continuously Adaptive Template Matching Algorithm for Human Tracking[C]//In 2017 First IEEE International Conference on Robotic Computing (IRC). Taichung: IEEE, 2017:303-309.

[12] HASSNER T, MASI I, KIM J, et al. Pooling Faces: Template Based Face Recognition with Pooled Face Images[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Las Vegas: IEEE, 2016:59-67.

[13] FRIDMAN L, LANGHANS P, LEE J, et al. Driver Gaze Estimation Without Using Eye Movement[J]. IEEE Intelligent Systems, 2015, 31(3):49-56.

[14] MURPHY C, ERIK T, MOHAN M. Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(2):300-311.

[15] HUANG C, DING X, FANG C. Head Pose Estimation Based on Random Forests for Multiclass Classification[C]// International Conference on Pattern Recognition (ICPR). Istanbul: IEEE, 2010:934-937.

[16] HUANG D, STOREY M, TORRE F D L, et al. Supervised Local Subspace Learning for Continuous Head Pose Estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs: IEEE, 2011:2921-2928.

[17] FOYTIK J, ASARI V K. A Two-Layer Framework for Piecewise Linear Manifold- Based Head Pose Estimation[J]. International Journal of Computer Vision, 2013, 101(2):270-287.