

PAPER • OPEN ACCESS

Region of Interest Coding Based on Convolutional Neural Network

To cite this article: Yang Jinkai *et al* 2021 *J. Phys.: Conf. Ser.* **1907** 012028

View the [article online](#) for updates and enhancements.

You may also like

- [Comparison of the scanning linear estimator \(SLE\) and ROI methods for quantitative SPECT imaging](#)
Arda Könik, Meredith Kupinski, P Hendrik Pretorius et al.
- [Comparison of region-of-interest-averaged and pixel-averaged analysis of DCE-MRI data based on simulations and pre-clinical experiments](#)
Dianning He, Marta Zamora, Aytekin Oto et al.
- [Automated PET-only quantification of amyloid deposition with adaptive template and empirically pre-defined ROI](#)
G Akamatsu, Y Ikari, A Ohnishi et al.



The Electrochemical Society
Advancing solid state & electrochemical science & technology

242nd ECS Meeting

Oct 9 – 13, 2022 • Atlanta, GA, US

Presenting more than 2,400
technical abstracts in 50 symposia



**ECS Plenary Lecture
featuring
M. Stanley Whittingham,**
Binghamton University
Nobel Laureate –
2019 Nobel Prize in Chemistry



Register now!



Region of Interest Coding Based on Convolutional Neural Network

Yang Jinkai^{1,a}, Wang Guozhong^{1,b*}, Zhu Liangqi^{1,c}

¹School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China

^aemail: m020318136@sues.edu.cn, ^cemail: lqzhu2020@163.com

^{*b}email: wanggz@sues.edu.cn

Abstract: The traditional region of interest coding method mainly uses low-level features to detect the Region of Interest (ROI). The ROI detected by it is poor in stability and is not easily interfered by noise. In this paper, ROI detection is performed on the image through a deep convolutional network to obtain a stable ROI based on the high-level feature extraction of the image, and then the discrete cosine transform (DCT) is performed on the image and divided into coding units. According to whether the coding unit is an ROI, To determine the quantization matrix used when encoding it. This article uses fine quantization for coding units that belong to ROI, and coarse quantization for non-ROI coding units. In this way, it can be ensured that the compression rate is greatly reduced without affecting the subjective perception of the image. Experiments show that the compression rate of this method can reach about 84%, and the weighted peak signal-to-noise ratio is improved by about 0.99dB on average compared with JPEG encoding.

1. Introduction

With the rapid development of image signal acquisition and display technology and the rapid development of the Internet, the penetration rate of high-definition video has greatly increased, 4k and 8k have also begun to spread, and the pressure on image transmission and storage is increasing. Although with the development of image coding technology in recent decades, the image compression rate has been greatly improved, but because the image resolution is getting higher and higher, the amount of compressed video data is still very large, so a more advanced method is needed. Efficient coding method. Taking into account the characteristics of the Human Visual System (HVS) ^[1], when humans are observing natural scenes, they often only focus on part of a scene, or consciously only observe a certain part of the scene. Specific scenery. We call this part of the focused observation area or specific area the area of interest. That is, the purpose of region-of-interest extraction is to locate the most sensitive and eye-catching region in the image. This paper proposes a coding based on the region of interest. For the region of interest, a fine compression with a low compression ratio is adopted, that is, more resources are allocated to the region of interest, sometimes even without compression, in order to obtain better results after decompression. Image effect. The background area of the image uses lossy compression with higher compression, that is, less resources are allocated to the background area, and sometimes the background area is not even transmitted. The purpose is to give priority to ensuring the quality of important information in the case of a larger compression ratio. This can greatly increase the compression rate without losing subjective experience.



2. Region of interest extraction

Region of interest extraction is the prerequisite of region of interest coding, and the quality of the ROI extracted from the region of interest directly affects the subjective experience of the decoded image. Traditional methods for extracting regions of interest usually use hand-made low-level features, such as color, texture, and contrast, to extract regions of interest. However, these methods that use underlying features to extract regions of interest in complex scenes have poor stability and are extremely susceptible to interference from image noise, and the actual effect is not good. In recent years, with the introduction and development of convolutional neural networks (CNN), convolutional neural networks can directly extract high-level, multi-scale semantic information from original images, and make many visual tasks have made great progress. Therefore, the extraction of the region of interest based on the advanced features has achieved a high improvement compared with the extraction of the region of interest based on the underlying features. There are many current region-of-interest detection algorithms based on deep learning. The most commonly used region-of-interest detection methods include DSS^[2], Amulet^[3], BDMP^[4], PiCANet^[5], etc. , although the detection accuracy of the region of interest Both have been greatly improved, but most of them are directly and indiscriminately applying multi-level convolution features. Due to the interference of redundant details, the results are not very good in terms of stability. However, PAGR^[6] uses a gradual attention-guided recurrent network to selectively integrate contextual text information from multi-level features, which can alleviate background interference and generate powerful attention features. Through the introduction of multi-path reflow connection, the use of global semantic information to guide the shallower feature learning process, essentially improving the entire network. Improved the stability and accuracy of the region of interest detection.

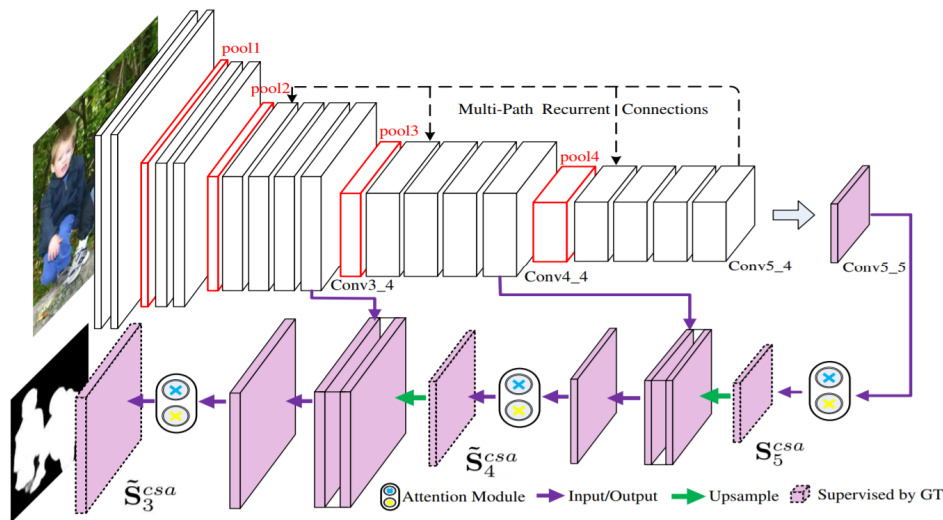


Figure 1 PAGR network structure

3. Region of interest coding

3.1. Two-dimensional DCT transform

The full name of DCT transform is Discrete Cosine Transform^[7], which is mainly used to compress some data or images. It converts the signal from the spatial domain to the frequency domain, because in this way, the relevant parts in the time domain can be separated in the frequency domain, so that the required frequency domain information can be retained in a targeted manner. Required frequency domain information. In fact, the DCT transform itself is lossless, that is, the original data can be restored losslessly in the inverse DCT transform. For the two-dimensional DCT transformation used for images, the mathematical formula is as follows:

$$F(u, v) = c(u)c(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i, j) \cos \left[\frac{(i+0.5)\pi}{N} u \right] \cos \left[\frac{(j+0.5)\pi}{N} v \right] \quad (1)$$

$$\text{where, } c(u) = \begin{cases} \sqrt{\frac{1}{N}}, u = 0 \\ \sqrt{\frac{2}{N}}, u \neq 0 \end{cases}$$

From the formula, we can see that if the two-dimensional image data is a square matrix, the formula satisfies:

$$F = AfA^T \quad (2)$$

$$\text{where, } A(i, j) = c(i) \cos \left[\frac{(j+0.5)\pi}{N} i \right]$$

Therefore, in practical applications, if it is not a square matrix, the data is generally filled and then transformed. After reconstruction, the filled part can be removed to obtain the original image information.

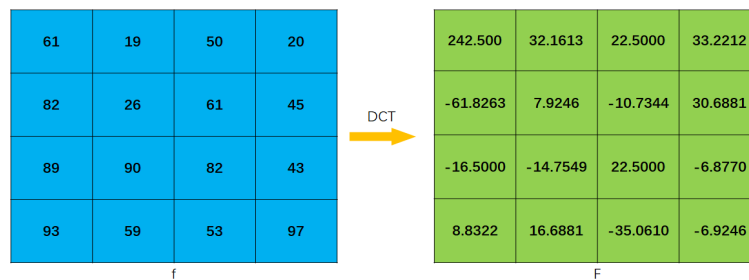


Figure 2 DCT transform

In fact, DCT cannot directly compress the image, but it has a good concentration effect on the energy of the image, laying the foundation for compression.

3.2. Two-dimensional inverse transform IDCT

IDCT transform is called Inverse Discrete Cosine Transform (Inverse Discrete Cosine Transform), DCT transform is to transform time domain information to frequency domain information, and IDCT transform is the inverse process of DCT transform, which transforms frequency domain information to time domain. The DCT transform has a very wide range of applications in the field of image analysis that has been compressed. Our commonly used JPEG still image coding and MJPEG and MPEG dynamic coding standards all use the DCT transform. The mathematical formula is as follows:

$$f(i, j) = c(u)c(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} F(u, v) \cos \left[\frac{(i+0.5)\pi}{N} u \right] \cos \left[\frac{(j+0.5)\pi}{N} v \right] \quad (3)$$

$$\text{where } c(u) = \begin{cases} \sqrt{\frac{1}{N}}, u = 0 \\ \sqrt{\frac{2}{N}}, u \neq 0 \end{cases}$$

Similarly, we can see from the formula that if the two-dimensional image data is a square matrix, the formula satisfies:

$$f = A^T F A \quad (4)$$

$$\text{where, } A(i, j) = c(i) \cos \left[\frac{(j+0.5)\pi}{N} i \right]$$

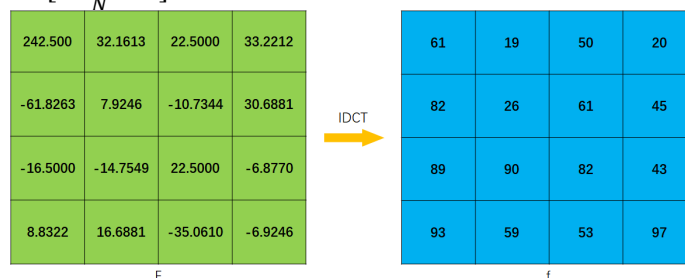


Figure 3 IDCT transform

3.3. ROI coding algorithm

Based on the combination of DCT transform and Huffman coding, we propose coding of interest. The idea is to divide the image into a region of interest and a background region (non-interest region). For the region of interest, we retain more frequency domain information, and for the background region, we retain only a small amount of frequency domain information, and then perform quantization and coding. In this way, the information we need can be retained as much as possible, and the subjective quality of the decoded picture is greatly improved. Specific steps are as follows:

- (1) Divide: Divide the picture into 8×8 sub-areas.

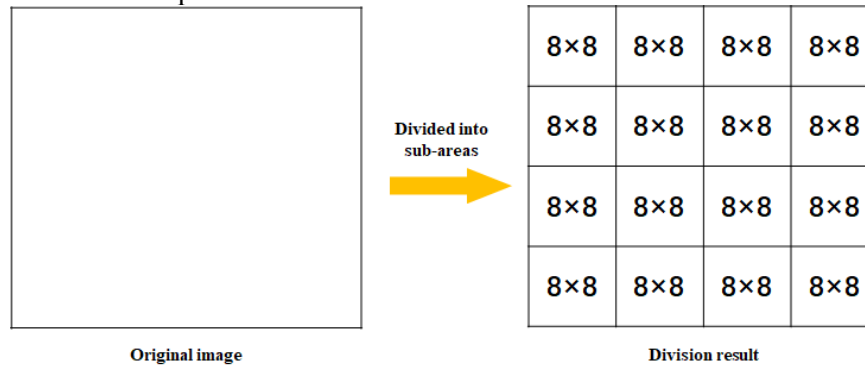


Figure 4 Subregion division of the image

- (2) DCT transformation: DCT transformation is performed on each sub-region.

		257.1	6.4	2.5	-0.3	0.4	0.1	-6.0	6.9
34	34	34	33	34	28	35	32		
34	34	34	33	34	28	35	32		
34	34	34	33	34	28	35	32		
34	34	34	33	34	28	35	32		
36	36	29	27	33	31	30	31		
32	32	35	30	32	33	31	27		
30	30	27	28	30	30	28	29		
8x8 sub-area		result							
		8.4	0.0	0.5	-5.0	1.9	3.4	-4.2	3.3
		-5.3	-1.0	-1.4	1.3	-0.7	-0.5	2.1	-1.7
		2.4	1.7	1.5	1.5	-0.6	-1.5	0.2	0.4
		-1.1	-1.6	-0.2	-1.8	1.6	1.2	-1.4	-0.1
		1.4	0.9	-1.9	-0.1	-2.0	0.9	1.5	0.7
		-2.0	-0.1	3.1	2.0	1.8	-2.7	-0.9	-1.3
		1.5	-0.2	-2.3	-1.9	-1.0	2.3	0.3	1.1

Figure 5 DCT transform coefficient

(3) Calculate the degree of coincidence with the region of interest: According to the degree of coincidence between the sub-region and the region of interest, calculate the degree of retention of frequency domain information. By discarding the high-frequency coefficients in each block, the purpose of image compression is achieved. Regions with a coincidence degree greater than $1/2$ retain more high-frequency coefficients, and for regions with a coincidence degree less than $1/2$, more high-frequency coefficients are discarded.

(4) Quantization: The quantization process is to divide the DCT coefficients by a certain quantization step size, and use different quantization precisions for the 64 DCT transform coefficients in an 8×8 DCT transform block to ensure that the specific DCT spatial frequency is contained as much as possible. Information, so that the quantization accuracy does not exceed the need. Among the DCT transform coefficients, low-frequency coefficients are more important to visual induction, so the quantization accuracy of the assignment is finer; high-frequency coefficients are less important to visual induction, and the quantization accuracy of the assignment is coarser. For different sub-regions, different quantization matrices are used for quantization. Even if $\gamma \cdot Q$ is used to quantify the image, the choice of γ depends on the result of step 3. The γ value used for the non-interest area is generally less than 1, and the γ value for the interest area is generally greater than 1. For example, when $\gamma=1$, the quantization matrix Q is used to quantize the DCT transform result.

$$Q = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

Figure 6 Quantitative matrix

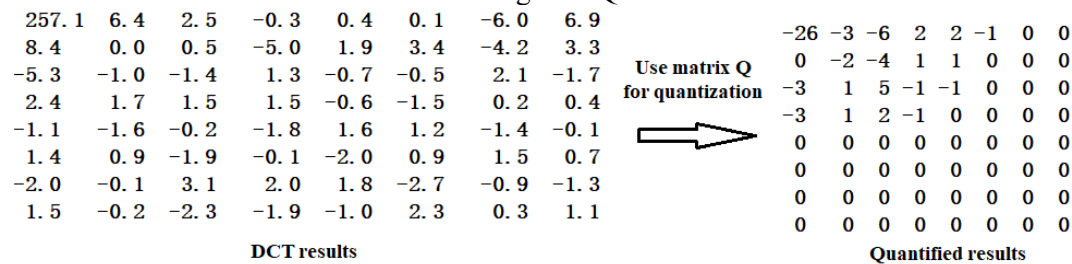


Figure 7 Quantization results of a subregion

(5) **Serialization:** In order to gather the non-zero numbers in the quantization result, serialization is performed in the direction shown in the figure below.

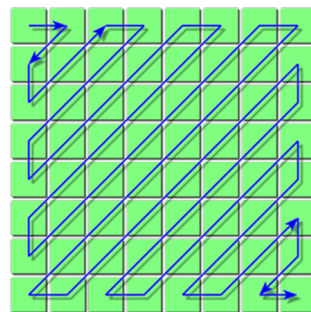


Figure 8 Serialization pattern

The serialization result is:

-26,-3,0,-3,-2,-6,2,-4,1,-3,0,1,5,,1,2,-1,1,-1,2,0, 0,0,0,0,-1,-1,0,0,0,0,. . . ,0,0

(6) Perform run length coding and Huffman coding on the serialized result, and finally get a sub-region sequence result.

91 CF FE A5 7F D1 BF CF FA 45

Generally speaking, when $\gamma < 1$, the sequence length is smaller than the sequence length when $\gamma = 1$, and when $\gamma > 1$, the sequence length is larger than the sequence length when $\gamma = 1$.

4. Evaluation method

Image quality evaluation is generally divided into subjective evaluation and objective evaluation. Among them, subjective evaluation relies on people's subjective feelings to evaluate merits, which is direct and convenient. It is in line with people's most visual and intuitive feelings. Objective evaluation is to establish some mathematical models based on some physical quantities of the image to evaluate the image quality according to some objective statistical data [8]. For example, the weighted peak signal-to-noise ratio of the image can be a good indicator of the quality of the image.

4.1. Peak signal-to-noise ratio (PSNR)

Peak of the Signal-to-Noise Ratio (Peak of the Signal-to-Noise Ratio)^[9] is a common parameter used to measure image quality. The peak signal-to-noise ratio is the larger the value, the better the image quality. Their expressions are as follows:

$$\text{PSNR} = 10\log_{10} \left[\frac{(2^n-1)^2}{\sum_{i=1}^M \sum_{j=1}^N [g(i, j) - \hat{g}(i, j)]^2} \right] \quad (5)$$

Among them, H and W are the height and width of the image respectively; n is the number of bits per pixel, generally taken as 8, that is, the number of pixel gray levels is 256; the unit of PSNR is dB, the larger the value, the smaller the distortion; $g(i, j)$ is the image before encoding, $\hat{g}(i, j)$ is the image after encoding.

4.2. Weighted peak signal-to-noise ratio (PSNR)

In order to better evaluate the quality of the coded image, according to the characteristics of the human eye, in an image, the definition of the area we care about is more important than the definition of the area that we don't care about. Therefore, by modifying the peak signal-to-noise ratio, the significance is obtained. Weighted signal-to-noise ratio (saliency region-weighted PSNR, SPSNR)

$$\text{SPSNR} = \alpha \times \text{PSNR}_{\text{ROI}} + (1 - \alpha) \times \text{PSNR}_{\text{non-ROI}} \quad (6)$$

Among them, PSNR_ROI is the PSNR value of the region of interest, PSNR_(non-ROI) is the PSNR value of the non-interest region, and the parameter α is the weighting coefficient of the saliency region, and its value range is 0.7-0.9.

5. Experimental results and analysis

In this chapter, we choose the ECSSD^[10] dataset for experiments, and select some typical pictures from it to illustrate our experimental results.

5.1. Subjective evaluation and analysis

As shown. First, we use the method in Chapter 4 to process image a) to obtain the saliency image (region of interest), which is image b). Then according to the method of this chapter, the region of interest coding and decoding (ROI coding and decoding) is carried out to figure c), and figure d) is the image obtained by using JPEG coding and decoding^[11].

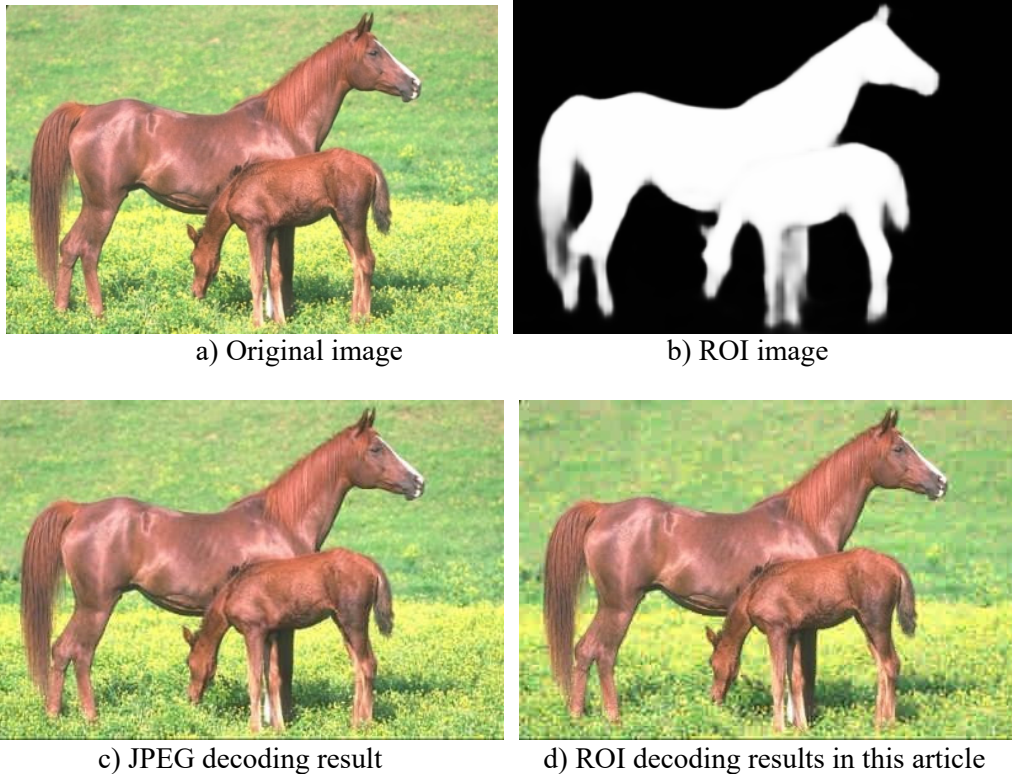


Figure 9 Comparison of ROI coding and JPEG coding in this article

Comparing figure c) and figure d), we can see that when we focus our eyes on the moment, the overall effect looks no different from figure d). When we look closely, we can see that the horse and background in figure c) are clear. Degrees are inconsistent. Among them, the horse's emotional degree is higher and the background clarity is lower. d) The clarity of the horse in the picture is the same as that of the background, the clarity of the background is higher than that of the picture c) the clarity of the horse is lower than that of the picture c). In order to observe more clearly, we show the details e) of Figure c) and Figure d). Observation shows that the details of Figure e) are more than those of Figure f). For example, there are more detailed images d) for the eyes of horses than images f). Therefore, this is in line with the characteristics of human observation, that is, when observing the image, I hope that the details of the information part I want are more.












Figure 10 The detailed comparison diagram of ROI encoding(right) and JPEG encoding(left)

5.2. Objective evaluation and analysis

We select three pictures with successively decreasing proportions of the region of interest for ROI encoding and decoding, as well as JPEG encoding and decoding at the same time, as a comparison.

Table 1 Coding results of images with different ROI ratios

Image name	Original image	ROI code	JPEG code
Horse			
Flower			
Person			

In order to show the advantages of ROI coding more intuitively, we respectively count the compression ratio and peak signal-to-noise ratio SPSNR of ROI coding and JPEG coding, as shown in

the following table:

Table 2 Comparison of compression ratio and SPSNR between ROI encoding and JPEG encoding.

Image name	Proportion of ROI	ROI encoding compression ratio	JPEG encoding compression ratio	ROI encoding SPSNR/dB	JPEG encoding SPSNR/dB
Horse	34.411%	80.10092%	77.6161%	30.76213	29.7008
Flower	13.374%	84.14448%	82.2621%	38.16242	37.8517
Person	02.957%	84.41425%	76.5708%	30.14806	28.5271

It can be seen from the table that the ROI coding method proposed in this chapter is better than JPEG coding in terms of compression ratio and weighted peak signal-to-noise ratio. And the smaller the proportion of the region of interest, the higher the compression of the ROI encoding, which is consistent with our subjective thinking.

6. Conclusion

This paper uses the convolutional neural network region of interest detection technology to realize the region of interest extraction and coding. According to the region of interest obtained by the convolutional neural network, we can perform different quantization coding according to whether the image block belongs to the region of interest. This ensures that the coding quality of the region of interest is consistent with the visual characteristics of the human eye. Without affecting the overall observation, the details of the area that people are paying attention to are more prominent. In addition, since most of the images are non-interest regions, the region-of-interest coding proposed in this paper greatly reduces the compression ratio. Of course, there are still some problems with the method in this paper, such as the extraction speed of the region of interest, and only the region of interest agreed by people can be detected. We will further study and improve in the follow-up work.

Acknowledgments

This article is one of the phased results of National Key R&D Program of China " (2019YFB1802700).

About the Author:

Yang Jinkai (1993-), Han nationality, male, from Bozhou, Anhui Province, master's degree student, main research direction is image processing, artificial intelligence.

E-mail: 459487218@qq. com, Tel: 15665321750, Address: 7707 Laboratory, 7th Floor, Building 7, Modern Transportation Engineering Center, Shanghai University of Engineering Science, Zip Code: 201620

Wang Guozhong (1962-), Han nationality, male, Shanghaiese, corresponding author, professor, doctoral tutor, main research direction is video coding and decoding, image processing, machine learning, etc. , E-mail: wanggz@sues. edu. cn

Zhu Liangqi (1994-), Han nationality, male, from Qingyang, Gansu Province, master's degree student, main research direction is image processing, artificial intelligence.

E-mail: lqzhu2020@163. com, Tel: 15651732732, Address: 7707 Laboratory, 7th Floor, Building 7, Modern Transportation Engineering Center, Shanghai University of Engineering Science, Zip Code: 201620

References

- [1] Bari A , Robbins T W . Inhibition and impulsivity: Behavioral and neural basis of response control[J]. Progress in Neurobiology, 2013, 108:44-79.
- [2] Hou Q , Cheng M M , Hu X , et al. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016:815-828.
- [3] Zhang P, Wang D, Lu H, et al. Amulet: Aggregating multi-level convolutional features for salient

- object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 202-211.
- [4] Zhang L , Dai J , Lu H , et al. A Bi-Directional Message Passing Model for Salient Object Detection[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018.
- [5] Liu N , Han J , Yang M H . PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection[J]. 2017.
- [6] Zhang X, Wang T, Qi J, et al. Progressive attention guided recurrent network for salient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 714-722.
- [7] Lu Yepin, Li Fengting, Chen Zhaolong, etc. Discrete Cosine Transform Coding Status and Development Research [J]. Journal on Communications, 2004(02): 106-118.
- [8] Chen Jianjian, Song Yuqing, Zhu Feng. Digital image processing and analysis[M]. Jiangsu University Press, 2015.
- [9] Yoo J C , Ahn C W . Image matching using peak signal-to-noise ratio-based occlusion detection[J]. Image Processing Iet, 2012, 6(5):483-495.
- [10] Shi J, Yan Q, Xu L, et al. Hierarchical image saliency detection on extended CSSD[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(4): 717-729.
- [11] Yin Guofu, Li Yunfei. Still image compression based on JPEG standard [J]. Science Technology and Engineering, 2008, 8(011): 3001-3003.