Color Compression of RGB Image Based on K-means Clustering

Mudan Lv[1,a], Xiao Wang[2], Yining Xue[3], Li Yu[4], Mingjie Jin[5]

[1]Nanchang Institute of Science & Technology, Nanchang, China

[2]Donghua University, Shanghai, China

[3]Beijing Forestry University, Beijing, China

[4]Dongbei University of Finance and Economics, Dalian, China

[5]Hangzhou Dianzi University, Dalian, China

[a]978745819@qq.com

Keywords: K-means clustering method; initial clustering center; color compression

Abstract: In life, the amount of information in an image is often very large. Color images usually contain thousands of colors, which brings great inconvenience to people's storage and transmission efficiency. Therefore, the image compression method is gradually being viewed by all. Under the premise of lossy compression, this paper focuses on the influence of RGB image before and after color compression on absolute mean error, establishes K-means clustering model, randomly selects the initial clustering center, and performs image color compression.

1. Introduction

The amount of information in an image is often very large, worth a thousand words. For example, the standard Lena test image has 148,272 colors. Such a huge number of colors has many adverse effects on the storage and transmission of images. There are two ways to address this adverse effect: lossy compression and lossless compression. This article only considers the lossy compression method of the image.

2. Problem Analysis

Given a color image with N colors, a new method is designed to reduce the number of colors to K (K < N), and the absolute average error of the image before and after the color reduction is minimized. Here, the K-means algorithm is first used to obtain a rough color compression result, and then aggregated according to a given criterion function until all colors are assigned to a specified number of groups, thereby implementing a color compression problem.

In model two, a new dynamic clustering method-mean-standard deviation clustering method is used to calculate the mean and standard deviation of the population first, then divide by aliquot. The idea selects the initial clustering center between the mean and the standard deviation, and then aggregates according to the given criterion function until all the colors are assigned to the specified number of groups, thereby implementing the color compression problem.

3. Related Work

3.1 K-means clustering[1]

The algorithm first selects K points from the sample as the cluster center, then calculates the distance of each object into the sample, and classifies the sample into the class of the cluster center closest to it. The average of each new cluster object is calculated as the new cluster center, and the iteration is repeated N times.

  DOI: 10.25236/mmmce.2019.147

3.2 Distance criterion

The distance criterion is the main basis and principle of the cluster analysis method. It mainly has the Euclidean distance and can be simply described as the geometric distance between the points in the multi-dimensional space. When the Euclidean distance between the points is less than a certain standard, Classify these points into the same class of objects.

4. Model Establishment

4.1 Model One

Step1: Sampling the signal of the picture, writing the coordinates of each color in the picture into a three-dimensional array, denoted as A(i, j.k), and extracting the total number of colors of any given image in the question.

Step2: Optionally select the initial cluster center[2]. For example, you can use the Random function to simulate and take k initial values. (k is the type of color that is compressed for the given final image).

Step3: Select the appropriate distance criterion for cluster analysis. In this paper, we choose the Euclidean distance[2], which can be defined as:

$$d(\vec{x},\vec{y})=\left[\sum_{p=1}^{m}\left|x_p-y_p\right|^2\right]^{1/2} \quad \square(1)$$

Where x, y are any two points in the image matrix, and m is the number of colors remaining except the initial cluster color number. The Euclidean distance between the various colors in the image and the initial clustering point can be obtained by the above formula. And thus determine the proximity of this color to its nearest initial color, the smaller the Euclidean distance, the more similar the two. The difference is smaller. This color can be grouped into the same class, also known as the same object.

Step4: Calculate the cluster center[3] of the new group (object) again. For each group, take the center of mass as the new initial cluster center.

Step5: Calculate the standard measure function until the program iterates N times, then stop, otherwise, continue steps 2 and 3.

- Solution of model one

The first question is solved by Matlab programming, and the Lena image is taken for testing. The number of colors of the compressed image is 64, 32, and 16 respectively to obtain the absolute average error. Table 1 shows the number of iterations of the program, and the color compression processing is 10 times. The results are shown in Figure 1 and Figure 2.

Table 1. K-means Algorithm for MAE

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
148272	64	11.6155
148272	32	15.0306
148272	16	20.3177

Figure 1 original & K=64

Figure 2 K=32 & K=16

- Model-one result analysis

The first question in this paper is the K-means algorithm, which is a classic algorithm for solving clustering problems. It is relatively simple and easy to accept. For the processing of large data sets, the algorithm maintains scalability and efficiency. The above K-means algorithm can see that the color processing results of the Lena image show that in the human visual range, when K=64 and K=32 are used, the two images are hard to detect and the original image. There is not much difference, so the algorithm can be used to roughly handle image color compression problems. However, it can be known from the data in Table 1: When the number of colors compressed at the end of the image is small, the absolute average error of the image with the original image is large, that is, the similarity is low. Therefore, this paper must seek a superiority. The method makes it possible to improve the similarity between the two graphs by optimizing the model and reducing the absolute average error between the two graphs when the number of compressed colors is the same. Therefore, this paper optimizes the results by changing the selection of the initial cluster center[4] and the determination of the distance criterion. quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)

4.2 Model Two

In the above analysis of the problem 1, the K-means algorithm is used in this paper. However, since the initial clustering center of the algorithm is randomly selected, the result of image compression may change accordingly. Therefore, it is very important to select the initial clustering center[5]. According to the characteristics of the data, most of the data are mainly distributed in the vicinity of the mean value of the middle body data. The absolute average error and the standard deviation are the indicators for judging the clustering effect. Therefore, in the second part of this paper, the mean-standard will be passed. The difference method is used to select the initial cluster center.

- Establishment of model one

Step1: Sampling the signal of the picture, and writing the coordinates of each color in the picture into a three-dimensional array, denoted as A(i, j.k), to extract the total number of colors of any given image in the question.

Step2: Selection of the initial clustering center [3]: First, find the population mean and the population standard deviation of the required image, which are respectively recorded as μ and δ. It is easy to know that most of the data are concentrated in the range of (μ- δ, μ+ δ). It is also assumed that the number of colors to be finally compressed is K, and (μ- δ, μ+ δ) is K-divided, and each halving point is taken as the initial cluster center. The i-th initial cluster center can be defined as:

$$m_i = (\mu - \delta) + \frac{2\delta_i}{K}, i = 1, \cdots, k \tag{2}$$

Step3: Select the appropriate distance criterion for cluster analysis. In model 2, the paper still uses

the Euclidean distance criterion, which is consistent with the calculation formula of the model, such as formula (1).

Step4: Calculate the cluster center of the new group again, and repeat step 2 to iterate continuously until the iteration is completed N times, and the clustering process ends.

Step5: Based on the iterated array, use Matlab to regenerate the new image and use the formula given in the problem to calculate the absolute mean error MAE between the original image and the compressed image.

- Solution of model two

In the second model, the Matlab program is used to solve the data mean and standard deviation, then repeat iteration, update the initial cluster center, get the best clustering object, and then use the program to find the compressed color image and the original image. Absolute average error value, the specific results are shown in the following Table 2: (This model assumes that the number of iterations is 10, and the Lena image is analyzed and compressed 64, 32, and 16 times respectively to compare with the model one).

Table 2. Model two for MAE

The number of colors in the original graphic	**Number of graphic colors after compressing color k**	**MAE**
148272	64	10.8923
148272	32	14.0547
148272	16	19.0652

The results of the Matlab program on Lena image compression are shown in Figure 3 and 4.

Figure 3 original & K=64

Figure 4 K=32 & K=16

- Model-two result analysis

From the solution process of the Matlab program in Table 2 above, it can be seen that when compared with the K-means algorithm in Model 1, when the number of iterations is 10, both take the Lena image for color compression test, when the final compressed color number is the same. The mean-standard deviation algorithm is slightly better than the K-means algorithm. For example, when K=64, the absolute average error of model one is 11.6665, and the absolute average error of model two is 10.68923, which is smaller than model one. Therefore, the mean-standard deviation algorithm has an optimization effect to some extent. However, by analyzing the above values, it can be found that the optimization effect of Model 2 is not obvious. The possible reason is still the selection of the initial cluster center. Therefore, in the model three, this paper selects the color with the highest

probability of occurrence as the initial cluster center to further the result. Improve.

4.3 Model comparison

The above are the two models established in this paper. In order to compare the advantages and disadvantages of the two models, this paper uses two models, Lena and Mandrills, to test. In the test process, the number of iterations is 10, and the model 1 is calculated separately. The absolute average error of the model 2 to the two images compressed to 64, 32, 16 colors, the results are shown in Table 3:

Table 3. Comparison Result

Model	**Lena(K_0 =148272)**			**Mandrill(K_0 =230426)**		
	K1=64	K2=32	K3=16	K1 =64	K2=32	K3=16
Model1	11.615	15.030	20.317	21.894	27.826	35.464
Model2	10.892	14.054	20.065	21.349	27.432	35.024
Model3	7.539	8.994	12.892	15.301	20.094	25.446

By analyzing the data in the above table, we can find that the model one is the traditional K-means clustering algorithm, and the absolute average error after compressing the image is the largest. The possible reason is that the algorithm is greatly affected by the initial value. Moreover, the results of clustering are often local optimal. Even if different initial values are chosen to repeat the algorithm, it is difficult to achieve global optimization. For model two, taking the same number of color compressions, the result is less than the absolute average error of model one, indicating that in model two, by introducing the mean-standard deviation to select the initial clustering center, blindness in model one can be avoided. Sexuality, however, data analysis shows that although Model 2 is better than Model 1, the effect of optimization is not obvious. The reason may be that the selection of the initial cluster center is not considered but the concentration of color distribution is concentrated.

In Table 3, take a K value. For example, when K=64, the absolute mean error of the K-means algorithm for compressing the Lena image is 11.6665, and the mean-standard deviation clustering method is 10.85923, and the probability clustering method is used. The result is only 7.5390. From this set of data, it can be seen that Arthur, which has the highest probability of occurrence, is selected as the initial clustering center to compress the graphics. The processing effect is better than other methods, and the Similarity between the compressed graphics and the prototype can be improved.

5. Model Test

5.1 Question Analysis

We use the method we designed to color-compact the image in material C and find the absolute average error value. Therefore, in this question, we choose to use the probability clustering model to test and substitute the given image.

Substituting the requested image into the model 2, using the Matlab program to solve the problem, the compression effect of each image can be obtained, and finally the absolute average error of each image and the original image is obtained. The result is as follows:

Table 4. Lena Image

The number of colors in the original graphic	**Number of graphic colors after compressing color k**	**MAE**
148272	64	8.9949
148272	128	7.0158
148272	256	5.8829

Compression result in Lean are shown in Figure 5 and 6.

Figure 5 original & K=256

Figure 6 K=32 & K=16

Table 5. Mandrill Image

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
230426	64	15.3017
230426	128	11.2558
230426	256	7.8623

Compression result in Mandril are shown in Figure 7 and 8.

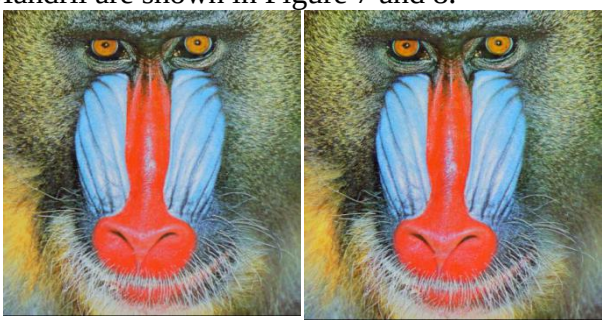

Figure 7 original & K=256

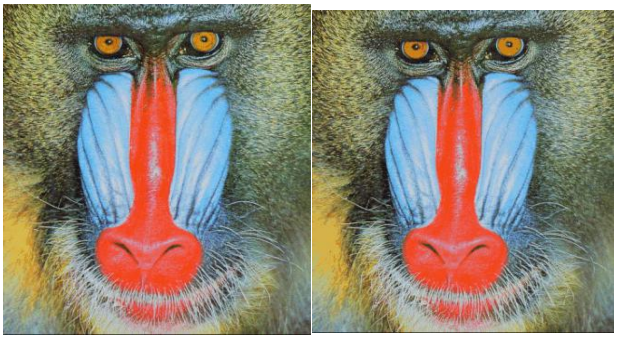

Figure 8 K=128 & K=64

Table 6. Parrots Image

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
71982	64	12.3914
71982	128	8.4361
71982	256	6.6169

Compression result in Parrots are shown in Figure 9 and 10.

Figure 9 original & K=256

Figure 10 K=128 & K=64

Table 7. Caps Image

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
34808	64	8.7076
34808	128	5.7455
34808	256	3.7976

Compression result in Caps are shown in Figure 11 and 12.

Figure 11 original & K=256

Figure 12 K=128 & K=64

Table 8. House Image

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
53312	64	10.7548
53312	128	7.5290
53312	256	5.4865

Compression result in house are shown in Figure 13 and 14.

Figure 13 original & K=256

Figure 14 K=128 & K=64

Table 9. Flowers Image

The number of colors in the original graphic	Number of graphic colors after compressing color k	MAE
37548	256	8.1050
37548	128	5.2087
37548	64	3.6513

Compression result in house are shown in Figure 13 and 14.

Figure 13 original & K=256

Figure 14 K=128 & K=64

In the second model of this paper, the problem of selecting the model-initial clustering center is simply optimized, and then the mean-standard deviation clustering method is used to select the initial clustering center, which effectively reduces the K-means clustering algorithm. The number of

iterations, speeding up the convergence of clusters, and each cluster can be more flat. The degree of polymerization. However, this method requires a comprehensive analysis of the distribution of the data and the calculation of the density of each data distribution before the model is built, and it brings a long running time to the multi-dimensional data participating in the clustering. Therefore, this paper will continue to introduce the probability clustering algorithm to select the initial clustering center, that is, use the model three pairs of model two to optimize.

References

[1] Banerjee A , Halder A . An efficient image compression algorithm for almost dual-color image based on K-means clustering, bit-map generation and RLE[C]// International Conference on Computer & Communication Technology. IEEE, 2010.

[2] Sun W , He Y . Spatial-chromatic clustering for color image compression[C]// IEEE World Congress on IEEE International Conference on Fuzzy Systems. IEEE Xplore, 1998.

[3] Cha J , Fausett L V . Comparison of three clustering algorithms and an application to color image compression [J]. Aerosense, 1997, 3077:225-235.

[4] Rogers S K. Comparison of three clustering algorithms and an application to color image compression[J]. Aerosense, 1997, 3077:225-235.

[5] Yu-Hai L U, Shen Y F, Wang C J, et al. Compression Algorithm for Computer Desktop Image Based on Color Clusering[J]. Computer Engineering, 2012, 38(21):221-225,236.