Design of Massive Image Duplicate Search Algorithm in Cloud Computing Environment

Shunyu Yao*
School of Management, Tianjin University of Technology, Tianjin, China

*Corresponding author e-mail: yaoshunyu@stud.tjut.edu.cn

Abstract. Massive image replay algorithm adopts cloud computing technology to realize image uploading, processing and classification on cloud platform. Image classification is realized by image classifier. Zernike matrix was used to compare the classified images and determine the pseudo-edge blocks around the image. According to the pixel value of the image, the binarization method was used to process the image. After processing, edge detection and line detection were carried out to determine the real image size. The similarity of two real images is calculated by normalized similarity criterion, and the image repetition is judged accordingly. The results show that the accuracy of the algorithm is higher than 99%, and the accuracy of the algorithm is higher than that of the comparison algorithm, and the accuracy of the algorithm is the highest when the similarity threshold is 0.8~0.9.

1. Introduction
Cloud computing technology is a distributed computing system composed of large-scale low-cost servers, which decomposes or classifies massive data through network cloud and transmits or feeds back the results to users. Cloud computing can provide cloud services based on user requirements, and has advantages such as low operating cost, high reliability, and good scalability. Cloud computing includes virtualization technology, distributed massive data storage, and distributed computing technology [1]. It can process massive and complex data in a short period of time and improve data processing efficiency. Image, as the main way of information transmission at present, is widely used in many fields [2]. How to judge the same image in a large number of images has become the main research content in the field of image replay at present.

2. Massive image duplicate checking algorithm in cloud computing environment

2.1. Massive image recognition technology in cloud computing environment
Mass image classification is the premise of image duplicate search. In order to better complete mass image classification, cloud computing technology is used to complete it. The whole process of image classification processing is realized on the cloud platform by using cloud computing technology [3]. Image classification requires first extracting image features from the image database on the cloud platform, matching the image features to be classified with the image features in the image database, and completing the classification of images according to the matching results. The image classification principle of cloud computing technology is shown in Figure 1. Among them, the main function of image
pretreatment is to complete the color conversion of the image and store the transformed image [4]. The method of correlation feature extraction is used to calculate the stored image data and obtain the image features [5].

![Figure 1. Principle of Cloud Computing Image Classification](image)

Image classifier is used to train image feature data samples, and the trained results are stored in local files for image classification. The main operation steps of classifier are described as follows: 1) Massive image data is uploaded on the cloud platform. After the upload and submission is complete, the data source is obtained from the distributed file system, the data is divided by data cluster configuration, the uploaded Reduce and Map are processed by classification, and node information in the Reduce and Map process is entered. 2) During operation, when the image samples stored in the distributed file system are read in, the combined parameters of the data samples after parameter type conversion are optimized by genetic algorithm, and the svM-train function is loaded in. To obtain the support vector, sample data training needs to be completed and the processing results are input into Reduce. 3) In the operation process of Reduce, the Map function of classification and sorting is converted into key/value in the form of data, the processed data is input to the specified path file, and the image classification result is output.

2.2. Image comparison based on Zernike matrix

2.2.1. Flow of Comparison Algorithm. Zernike matrix is used to complete the comparison of classified images, and the steps are as follows: 1) Since there may be false edge blocks after image rotation, in order to remove them, false edge block detection of two contrast images (image A and B) should be completed first to determine the region and size of the two images. After the removal of false edge blocks, the effective contents of the image region can be saved, namely A1 and B1. 2) The interpolation method is used to normalize the image of B1, so that the size of B1 and A1 is the same, and B2 is obtained. 3) Rotate A1, rotation times is S, each rotation Angle is 360/s, calculate the TTH Zernike moment of A1 after each rotation, and T ≥ 2, construct s *T matrix according to the calculated value, where is
4) Calculate the mean and standard deviation of every column of the matrix $K_A$, get the mean vector and standard deviation vector.

\[
\bar{K}_t = \frac{1}{T} \sum_{s=1}^{S} K_{st}, n = 1, \ldots, T
\]

\[
D_t = \left( \frac{1}{T-1} \sum_{s=1}^{S} (K_{st} - \bar{K}_t)^2 \right)^{\frac{1}{2}}, t = 1, \ldots, T
\]

5) Without rotation of B2, T Zernike moments corresponding to A1 are calculated and the moment vector $V_B$ is obtained. 6) According to the $K_{avg}$, $D_{div}$ and $V_B$ of the obtained image A1, we can judge whether the two images are the same. The specific process is shown in Figure 2.

---

**Figure 2.** Process of Algorithm Comparison

---

2.2.2. Determination of real image area size. After the image is rotated, there will be pseudo-edge blocks around, usually the area of the pseudo-edge block is black, white or other single gray area. In order to solve the influence of pseudo-edge blocks on Zernike moment value, all pseudo-edge blocks should be distinguished. After scanning the pixel values near the four edges of the image, the occurrence probability of the pixel values is calculated. If the proportion of a pixel value is large, it is judged that the pixel value will form a pseudo-edge block pixel value after rotation, and the connected image area adjacent to the four edges of the pixel value is judged as a pseudo-edge block. In order to determine the actual size of the real image and ensure subsequent image size normalization, the processing steps are as follows: 1) In order to transform the image into black and white image, binarization method is adopted to complete the image processing according to the pixel value of the image. The pseudo-edge block area was regarded as a separate category, and the other areas were classified as another category, and all were judged to be real images. 2) The binary black and white image is detected by edge detection as well as the existing lines, and the four lines intersecting in the image which can form a rectangle are regarded as the boundary of the real image. 3) Determine the rectangle formed by the intersection of four straight
lines, regard it as the real image area, and determine the real image size according to the coordinates of the intersection of four straight lines.

2.2.3. Normalized similarity criterion. The similarity between real images is measured by similarity criterion, and its value range is between [0, 1]. When the order of Zernike moment is high, there is a great difference between the calculated result and the moment value of lower order, even there are several orders of magnitude difference. In order to ensure the uniformity and equalization of each Zernike matrix, \( V_B \) is normalized, \( V_B^N = [V_{B2}^N, V_{B3}^N, ..., V_{BT}^N] \), where is

\[
V_{Bt}^N = \frac{V_{Bt}}{K_t}, t = 1, ..., T
\]  

(4)

According to the numerical results obtained from the calculation of each element in \( V_B^N \), the similarity judgment of real images is completed. If the numerical results are close to 1, the two images are relatively similar. The similarity of \( \bar{K}_{avg} \) and \( V_B \) is calculated. If \( \bar{K}_{avg} \) has been normalized and become a full 1 vector, the similarity of \( \bar{K}_{avg} \) can be as follows:

\[
K = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{V_{Bt}^N}{\sqrt{\sum_{t=1}^{T} (V_{Bt}^N)^2}}
\]  

(5)

If the obtained similarity \( \kappa \) value is greater than the set threshold, it indicates that the two images are identical; otherwise, they are not. According to the similar results, the image duplication is judged and the image duplication is completed.

3. Simulation test results and analysis
Select an image library as the research object, carry out the relevant test analysis. There are 82000 images and 10087 duplicate images in the library. There were 14,600 landscape images and 4,220 repeated images; 12,800 architectural images and 3,020 human images; 18400 text images and 1120 repeated images; 11200 animal images, 728 repeated images; There were 25,000 toy images and 999 duplicate images.

3.1. Classification performance testing
The image classification performance of the algorithm in this paper was tested from two aspects: the influence of node number on image recognition time and image classification accuracy. The test results are shown in Table 1 and Table 2. By analyzing Table 1, it can be seen that in the process of image recognition by the algorithm in this paper, if there are only two nodes on the cloud computing platform, it takes a long time to exchange toy image data. This phenomenon indicates that for image recognition, the time required by two computers is larger than the time required by one computer. When the number of nodes is more than 3, the time required to process the same number of images gradually decreases with the increase of the number of nodes. The test results show that the increase in the number of nodes will increase the speed of image classification, and the appropriate number of nodes can be selected according to the number of images to be classified.

| Table 1. Image recognition time under different number of nodes/ms |
|--------------------------|-------|-------|-------|-------|
| Number of toy images/piece | 1  | 2  | 3  | 4  |
| 5000  | 75  | 73  | 71  | 69  |
| 10000 | 181 | 186 | 166 | 134 |
| 15000 | 374 | 440 | 301 | 263 |
| 20000 | 702 | 704 | 612 | 580 |
| 25000 | 911 | 1411| 822 | 721 |
Table 2 shows that: For five kinds of images, the algorithm of classification accuracy, the best classification accuracy is above 99% comparison algorithm is higher than the number of leakage and a deterrent to algorithm in this paper, the classification accuracy is low, this is mainly because the algorithm using cloud computing technology from through the distributed file system access to huge amounts of image data, through the cluster configuration data processing data, Ensure the accuracy of image classification results.

| Image category   | Test index               | Algorithm | Algorithm1 | Algorithm2 |
|------------------|--------------------------|-----------|------------|------------|
|                  | Misidentification quantity/piece | 28        | 388        | 401        |
| Landscape image  | Leakage quantity/piece    | 32        | 209        | 198        |
|                  | nicety of grading/%       | 99.59     | 95.91      | 95.88      |
| Architectural image | Misidentification quantity/piece | 33        | 432        | 392        |
|                  | Leakage quantity/piece    | 23        | 271        | 364        |
|                  | nicety of grading/%       | 99.56     | 94.51      | 94.09      |
| Human image      | Misidentification quantity/piece | 30        | 388        | 404        |
|                  | Leakage quantity/piece    | 34        | 322        | 312        |
|                  | nicety of grading/%       | 97.88     | 96.14      | 96.11      |
| Text image       | Misidentification quantity/piece | 45        | 441        | 296        |
|                  | Leakage quantity/piece    | 32        | 302        | 456        |
|                  | nicety of grading/%       | 99.58     | 93.34      | 93.29      |
| Toy image        | Misidentification quantity/piece | 44        | 586        | 601        |
|                  | Leakage quantity/piece    | 32        | 387        | 370        |
|                  | nicety of grading/%       | 99.70     | 96.11      | 96.12      |

3.2. Perform a performance check test
To further test the replay lookup performance of the algorithm in this paper, a group of human image images are randomly selected, as shown in Fig. 3(a) is the original given image. After 60% scaling, (b) is obtained by counterclockwise rotation. At this time, the contents of the two images are the same, but the data itself is quite different. The fourth-order Zernike moment values (including 9 Zernike moment values in total) were selected to rotate the figure (a) at an Angle of 20° each time to obtain the Zernike moment value and standard difference value of figure (a) and the Zernike moment value and normalized value of Figure (b), as shown in Table 3 and Table 4 respectively. By analyzing Table 3, it can be seen that the table contains the mean value and standard deviation of Fig. 3(a). Compared with the mean value, it can be seen that the standard deviation is relatively small, indicating that the size of Zernike moment remains relatively stable when the image is rotated at different angles, indicating that the algorithm in this paper has good rotation invariability. By analyzing Table 4, it can be seen that the obtained moment value is calculated by formula (5) to obtain the similarity value. The higher the similarity value is, the content of the two images is the same, indicating that the two images are repeated. The method presented in this paper has the ability of image duplicate checking and can complete the duplicate checking of massive images.

![Figure 3. Experimental Images](a) (b)
Table 3. The Mean & Standard Deviation of Fig 3. (a)

| Zernike Moment Order | Mean  | SD  |
|----------------------|-------|-----|
| Z00                  | 38.14 | 0.10|
| Z11                  | 107.86| 0.27|
| Z21                  | 343.12| 0.79|
| Z23                  | 228.78| 0.58|
| Z30                  | 1294.17| 3.17|
| Z32                  | 431.37| 1.08|
| Z41                  | 654120.06| 1594.61|
| Z43                  | 143370.72| 348.73|
| Z45                  | 762.6 | 1.88|

Table 4. The Moment Values & Normalized Values of Fig 3. (b)

| Zernike Moment Order | Moment Values | Normalized Values |
|----------------------|---------------|-------------------|
| Z00                  | 38.12         | 0.99949           |
| Z11                  | 107.79        | 0.99938           |
| Z21                  | 342.96        | 0.99955           |
| Z23                  | 228.64        | 0.9994            |
| Z30                  | 1293.35       | 0.99938           |
| Z32                  | 431.12        | 0.00042           |
| Z41                  | 653699.85     | 0.00037           |
| Z43                  | 143276.69     | 0.99935           |
| Z45                  | 762.12        | 0.99939           |

Image replay can be understood as clustering repeated images into the same cluster. Therefore, the measurement formula of replay effect is as follows:

\[
acc(A) = \left( \sum_{a \in A} \max - \text{dup}(a) \right) / N
\]

Where is: A is the result set of image repeated detection, and its elements are detected repeated images. The maximum number of true repeated images in A is max-Dup function. If \( a = [1, 1, 2, 2, 2, 3, 3] \), max-dup \((a) = 3\), the number of 2 in table 2 is the number of elements that appear most in A, acc means to check the purity of reunion class.

Taking the character image data set as an example, three algorithms are used to query its similarity. Under the condition that the similarity threshold of the three algorithms changes, the change results of ACC are shown in Figure 4. The analysis of Figure 4 shows that the ACC value of the algorithm in this paper is higher than that of the two comparison algorithms when the similarity threshold is changed, indicating that the algorithm in this paper has the best effect of image similarity check. The ACC value of the two comparison algorithms is relatively low, and a large number of images are incorrectly detected as duplications due to the change of threshold value. The proposed method has good classification performance and can divide images of the same class into a set, which greatly reduces the number of errors in image similarity detection and ensures the accuracy of image similarity detection. In addition, according to the curve changes in the figure and the fixed range of threshold, it can be seen that acc accuracy of the algorithm in this paper is the highest within the range of similarity threshold of 0.8–0.9.
4. Conclusion
In order to realize the duplicate checking of mass image content, a duplicate checking algorithm of mass image is designed under the cloud computing environment. The test results show that the number of nodes in cloud computing technology has an impact on mass image classification, and appropriate nodes can be selected according to the number of images. For the five types of images, the classification accuracy of the algorithm in this paper is high, which lays a reliable foundation for the high-precision retrieval of subsequent images. The proposed algorithm has good rotation invariance and can effectively complete image duplicate checking.

References
[1] Wang, L. and Sun, H.Y., Image Retrieval Based on Improved Visual Descriptors in Cloud Computing, J. Control Engineering of China, 27(2020) 554-560.
[2] Yang, X.H., Yang, J. et al., Improved TLBO Algorithm for Image Edge Detection in Cloud Environment, J. Science of Surveying and Mapping, 42(2017) 188-194.
[3] Xing, W.K., Gao, X. X.et al., Research on Fuzzy Decoupling Energy Efficiency Optimization Algorithm in Cloud Computing Environment, J. Computer Science, 44(2017) 75-79.
[4] Liu, X.J., Yuan, J.B. et al., Data Distribution-oriented K-means Clustering Algorithm in Cloud Computing Environment, J. Journal of Chinese Computer Systems, 38(2017) 712-715.
[5] Xu, J.R. and Zhu, H.J., Multi-objective Scheduling Algorithm for DAG Task in Cloud Computing Environment, J. Application Research of Computers, 36(2019) 31-36.