IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-11, Nov 2019 E-ISSN: 2347-2693

A Multi Factor Duplicate Image Deduplication System

Rounak A. Samdadia[1*], Nitin Patil[2]

[1] Department of Technology, Savitribai Phule Pune University, Pune, India
[2] Department of Computer Science, Savitribai Phule Pune University, Pune, India

[*]*Corresponding Author: rounakajs@gmail.com, Tel.: +91-95951-95037*

DOI: https://doi.org/10.26438/ijcse/v7i11.4951 | Available online at: www.ijcseonline.org

Accepted: 11/Nov/2019, Published: 30/Nov/2019

***Abstract*—** With the invent of internet technology and increasing the use of digital technology, the storage system has been into a difficult mode of expansion as the growing data is a big concern for the storage system, typically the images which occupy large scale size of storage. Hence worldwide data Deduplication mechanism had been used enormously to improve the backup storage system.

Lots of efforts had done worldwide for identifying the image Deduplication system to improve the storage utilization system. In spite of these continuous efforts the so-called existing system can only manage to remove the images which are the same in texture and size, but it fails to identify the Deduplication of images that may have the same visual perceptions but may have different effects. To solve the above issue of image Deduplication our thesis work proposed the mechanism to identify the duplicated images which are technically the same from the perspective of visual identification but different with some effects. With further improving to remove such duplicate images, we are contributing to improving storage space. Our proposed algorithm based on feature selection and extraction and accuracy optimization will not improve storage space but helps to perform such task quite efficiently and effectively.

***Keywords*—** Image Deduplication, Feature Extraction, Centroid selection, Storage space, Optimization, DHash, Hamming, SIFT Algorithm

I. INTRODUCTION

With the increasing use of web technology and widespread use of digital technology the amount of data on the digital world is increasing with the exponential pace at an alarming rate. It was first identified in the year 2007 that the total volume of digital data components had crossed the global storage capacity and it is also observed that by the year 2011 only ½ of digital content can be stored globally. With this, it is highly impossible to make a solution to this digital data explosion by increasing the storage devices' facilities through data centers.

Data Deduplication is a method that ensures only one unique instance of data is present in the database. Deduplication has been widely used in cloud and big server systems to improve the storage utilization and effective usage of storage systems. Similarly, eliminating the redundant images from the storage or system can be stated as Image Deduplication.

Our module is designed to provide on the go solution, so hashing is done on the sender's side itself. The storage optimization technique is more powerful. It will be an image sharing kind of module. So, it can be used to share images in a large amount too. We intended to build it for the local database, which will ultimately save lots of processing and space on our device.

The rest of the paper is organized as follows: In the next section II, we will provide the related work as a literature review. In Section III, we will provide our system development approach. Section IV consists of results and discussions, and finally, we will conclude with future scope in section V.

II. RELATED WORK

In the close to duplicate image recognition paper authored by V. B. Nemirovskiy, Tomsk, and Usage of multi-step segmentation for close to duplicate image recognition is investigated.

The search pattern supported the rank distributions of the brightness clusters cardinality is usually recommended. Experimental results on the near-duplicate image recognition supported the applying of the steered search pattern are given.

It is shown that the utilization of a multi-step segmentation and rank distributions of the brightness clusters cardinality

permits to with success recognizing the duplicates, that received by the considerable.

Dynamic Data Deduplication in Cloud Storage by Waraporn Leesakul, Paul Townend, Jie Xu states that Cloud storage is one of the services provided in cloud computing which has been increasing in popularity. The main advantage of exploitation cloud storage from the customers' purpose of view is that customers will scale back their expenditure in getting and maintaining storage infrastructure whereas solely paying for the quantity of storage requested, which might be scaled-up and down upon demand.

Data Deduplication techniques are dropped at improves storage potency in cloud storage. With the dynamic nature of knowledge in cloud storage, information used in the cloud changes over time, some information chunks could also be scan oftentimes in the amount of your time, however, it might not.

A study on data Deduplication techniques for optimized storage by E. Manogar, S. Abirami In recent years, the explosion of the data such as text, image, audio, video, data centers, and backup data lead to a lot of problem in both storage and retrieval process. There are 2 existing techniques for eliminating redundant information within the storage system like information Deduplication and information reduction.

Data Deduplication is one in every of a method that eliminates redundant information, reduces the information measure and additionally minimizes the disk usage and value.

This paper tries to summarize numerous storage optimization techniques, ideas, and classes mistreatment information Deduplication.

In addition to the present, chunk based mostly information Deduplication techniques area unit surveyed well.

III. SYSTEM ARCHITECTURE AND IMPLEMENTED APPROACH

This section provides the implementation mechanism we had used in our work using improved DHash Mechanism and Hamming distance calculation.

Initially, we had converted the source image into a greyscale format and also converted the images dataset into the greyscale format. Then we had resized and flattened these images to one common size format so that comparison will be made easy. After that, we had calculated the intensity difference. In this step, we had calculated the gradient intensity difference which is nothing but the difference between rows and columns values. Then we had compared the original image with every image in directory, created the array and find out the hamming distance using the hamming distance formula. For eg the original images array hamming is [1,0,0,0,0] and image to compare array is [0,0,0,0,0] Then hamming distance is 1/5=0.20. If the hamming distance after comparison is less than 0.20 then we will be adding to the dictionary data structure which we had created. Then towards the end, we will be checking all the entries from the dictionary data structure and displaying all the duplicated images from the source directory.

.

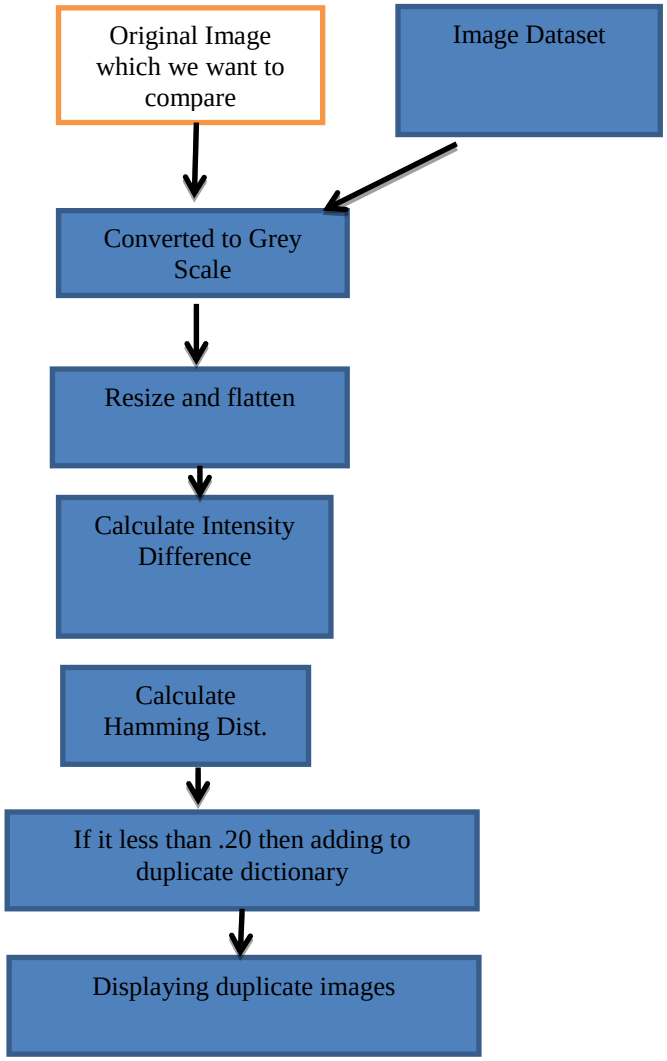

Figure 1: Flowchart of Modified DHash Algorithm

IV. RESULTS AND DISCUSSION

We had tested and tried two algorithms for identifying the similarity between two images which are changed with different effects. We had almost change the original image with 11 effects which include grey scaling, blur, blue effects and so on. The two algorithms which we had tested are

DHash and hamming techniques based on MD5 and another is SIFT algorithm of OpenCV image processing libraries. What we had found that with the help of a hamming distance along with the DHash technique, our proposed algorithm performs better than SIFT Algorithm in terms of finding out duplicate images with different effects and sizes.

Conclusion and Future Scope

In this study, for the experimental results we had compared original image with 12 different effects like sharp, sunburst, blue, blur and so on and when we compared the duplication effects its shows that the planned approach can do higher Deduplication rate and Deduplication accuracy by setting appropriate threshold value of T which is 0.20 in terms of hamming distance. Our algorithm had proven far better results than existing dHash and SIFT algorithm

In the future, we will further expand the definition of duplicate images and use a new algorithm to identify a variety of image transformations such as rotation, shift, watermark and so on. Moreover, future work also focuses on a new index structure to further improve Deduplication speed and scalability.

We can also focus to implement the system with Machine learning mechanism in the future where the data size will be very big and we can do the prediction using ML algorithm for the betterment of various devices for optimizing the space complexity.

REFERENCES

[1] NEMIROVSKIY V.B., STOYANOV A.K. Near duplicate image recognition based on the rank distribution of the brightness clusters cardinality, COMPUTER OPTICS. – 2014. – VOL. 38(4). – P. 811-817.

[2] Dynamic Data Deduplication in Cloud Storage by Waraporn Leesakul, Paul Townend, Jie Xu. IEEE 8th International Symposium on Service-Oriented System Engineering (SOSE 2014). IEEE 8th International Symposium on Service-Oriented System Engineering (SOSE 2014), 07-11 Apr 2014, Oxford, UK. IEEE, pp. 320-325. ISBN 9781479925049

[3] A study on data Deduplication techniques for optimized storage by E. Manogar, S. Abirami, 2014 Sixth International Conference on Advanced Computing(lCoAC),pp 161-165, 978-1-4 799-8159-5114

[4] Eunji Lee, Jee E. Jang, Taeseok Kim, Hyokyung Bahn, "On-Demand Snapshot: An Efficient Versioning File System for Phase-Change Memory," IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 12, December 2013.

[5] Q. He, Z. Li, X. Zhang, "Data deduplication techniques, "Future Information Technology and Management Engineering (FITME)," vol. I, pp. 430-433, 2010

[6] Maddodi.S, Attigeri G.V, Karunakar. A.K, "Data Deduplication Techniques and Analysis," Emerging Trends in Engineering and Technology (lCETET), pp 664 - 668, IEEE, 2010

[7] N. Mandagere, P. Zhou, M.A. Smith, and S. Uttamchandani."Demystifying data deduplication," In Proceedings of the ACM/IFlP/USENIX Middleware'08 Conference Companion, pages 12-17. ACM, 2008

Authors Profile

Rounak Samdadia: Born in Pune on August 24, 1991. He completed his Bachelors in Engineering (Computers) in 2014 from VIIT, Pune under Savitribai Phule Pune University, Pune, Maharashtra, India. This paper was written in regards to the research done under the education of Masters in Technology (Computers and Information Technology) in 2019 from DoT, under Savitribai Phule Pune University, Pune, Maharashtra, India with majors in Image Processing. His research domains include image processing and face recognition.

Nitin P. Patil is an Assistant Professor at the Department of Computer Science in the Faculty of Science and Technology at Savitribai Phule Pune University. He earned both his bachelor's degree and a master's degree from North Maharashtra University Jalgaon in Computer Science. Currently, his research involves in the field of Text Mining, Image Processing, Content-Based Image/Information Retrieval.

