IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 4, April 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104077 |

Survey on Duplicate Image Finder

Namrata Gaikwad, Sahil Sapkal, Pratykash Rai, Prof. Kanchan Doke

Department of Computer Science, Bharti Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Department of Computer Science, Bharti Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Department of Computer Science, Bharti Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Department of Computer Science, Bharti Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

ABSTRACT:- Duplication of images is a common issue faced by mobile users. Nowadays everyone uses a smartphone. A huge number of similar data is shared on the internet mobile users receive several duplicated images, files. This leads to a lack of storage, reducing mobile phone performance. Detecting such replicate images, files are challenging due to the large collection of images, files present in mobile storage, if we try to delete such similar images it is time-consuming and requires a lot of effort. This paper presents a survey on different methods and techniques used for the detection and elimination of replicate images and files. To overcome this replicate images and file problem we will be developing a Duplicate Image Finderapplication.The proposed system is divided into three modules These modules are Duplicate images, pdf, audio detection.The proposed system focuses on the deletion of similar images, pdf, audio files in one click

KEYWORDS – Duplicate, image, detection, Base64

I. INTRODUCTION

The most important task for the mobile user is to keep mobile devices running as high as possible. Lack of storage space is the most important issue.As there isa huge number of replicate image files one cannot delete them manually. The process of manually finding and deleting such kinds of duplicate images and files consumesa lot of time.Due to such unwanted images and files, phone memory slows down.With the rapid usage of the Internet, many socialnetworking apps suchas Facebook, WhatsApp, Instagram are used for communication and sharing of images, videos, audio, files. So the mobile users have several images and files from these social media apps. As mobile users receivelots ofsimilar and unwanted images, files. Due to this, the mobile device storage space is occupied by such type of replicate data. The duplicate images are of two types exact duplicate and near-duplicatestheexact similar images are 100% similar and the near-duplicate image has some minor changes this type of imageis rotated,cropped, or edited.The same image with a different version is called near-duplicate. This type of replicate image and file must be deleted to increase device efficiency and performance. To find Similarities between images is a challenging task. Many techniques are introduced to find such identical images some of these techniques have their pros and cons various techniques are surveyed in this paper.
The survey paper is organized as follows.

II. RELATED WORK

The paper explained below is taken for the study of the survey conducted. Each paper explains different methods and techniques which are as follows.

Y. Ke, Rahul Suktankar [1], Larry Huston provides a system closer to the duplication of detection and extraction of image. In the nearest duplicate detection, the problem is resolved using an effective search for steady detection points (Dog Detection), local descriptors (PCAsift), and restrained data (LSH). It uses a unique local descriptor to provide several conversions and high-quality matches based on spare parts based onpresentations of images. Hashing with location is used for local descriptor indexes. Hashing helps to get an efficient data layout for interactive operations. The drawback of using a part-based approach has to query hundred to thousand of features at an instance which could make it slow. While our experiments show glorious ends up in retrieving nearduplicates and sub-images, there ar limitations to our technique. Oursystem will match similar pictures of a similar scene notwithstanding they were not near-duplicates nor cast pictures, as in our MM270K database. Thisscenario may occur if the proprietary image info contains photos of famed landmarks, wherever there ar seemingly to be several photos of a similar landmarks on the

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104077 |

online. Our system may notice similar keypoints on the landmarks and incorrectly match them. Others have exploited this property as a feature to notice images of a similar scene taken from completely different camera locations as in [17, 18, 20]. To informally take a look at the performance of our system under such conditions, we tend to gathered fifty nine photos from the online of Big Ben, Eiffel Tower, and 0.5 Dome, and queried them against eighteen professional photos of a similar landmarks. we tend to were happy to see only 1 match (i.e., false positive) among the fifty nine queries. The reason why we tend to don't accidentally establish similar scenes as cast images additional oftentimes is thanks to well-known limitations in applying current interest purpose detectors to acknowledge real-world objects.

K.K.Thyagharajan G.Kalaiarasi[2], presents approaches and features extraction methods for near-duplicate images the techniques and methods involved in this are pixel extract method, Keypoint-based extraction, clustering methods. Drawbacks involved in pixel-based methods - Complexity in NDD, Computational complexity. The search engines face a drag resulting in the very fact thatthe search results square measure of less relevancy to the user thanks to thepresence of near-duplicate pictures. Since the digital contentis widespread and conjointly simply redistributable, the presence ofnear duplicates affects the performance of the search enginescritically. during this paper, the progressive approaches to detecting near-duplicate pictures and their applications square measurereviewed besides the methodologies used. The mainchallenges during this feld square measure mentioned. This review providesresearch directions to the guy analysisers WHO have an interest to figure during this feld. Also, the survey targeted solelyon the feature extraction task in laptop vision system.There square measure more tasks and techniques to try to to analysis.

Ammar Asaad, Ali Adil Yassin Alamri[3] due to duplicate files low storage space of the device is the main issue this paper shows how to overcome this issue by using Huffman code and hash function MD5.Huffman code generates unique code for each image and save this code to the image in the reference file and use it to extract duplicate in the device storage. The proposed system removes the duplicates as a result the device work in the best way.One of the most important challenges for smartphone users is to keep the device working as high as possible. Low storage space is the most important issue. We solved this issue by the used hash function (MD5), Huffman code to generatea unique code for each image and save this code in the image in the index file and use it to rid of a duplicate from the device's storage. Removing images that have the same code and maintaining one copy. The proposed scheme get a good result in removingthese duplicate images fast. We applied it to two devices (Galaxy A5 [2017] SM-A520F, HUAWEI CUN-U29). As a final result, the performance of the device's CPU is better thanmaking the device work in the best way. In future work, in relation to the image matching schemes, the user may add something to the image, and as a result, a new image is obtained, but it is still the same image with a simple change. In future work, we will try to solve this issue.

One of the most important challenges for smartphone users is to keep the device working as high as possible. Low storage space is the most important issue. We solved this issue by the used hash function (MD5), Huffman code to generating unique code for each image and save this code in image in the index file and use it to rid of a duplicate from device's storage. Removing images have the same code and maintain one copy. The proposed scheme get a good result in remove this duplicate images in fast. We applied it at two devices (Galaxy A5 [2017] SM-A520F, HUAWEI CUN-U29). As a final result, the performance of device's CPU is better that makes the device work in the best way. As future work, in relation to the image matching schemes, the user may add something to the image, and as a result, new image obtained, but it still the same image with simple change. In future work, we will try to solve this issue.

Somchai, Wen Dong[4], presented the Base64 Encoding explanation in detail.Base64 Base64 is a binary to text-to-text encoding system. Represents binary data in printable ASCII alphabet unit format by translating it into a radix-64 representation. Base64 encryption is commonly used when it is necessary to transfer binary data through binary data encryption and is designed to address the textual data that is part of the 7-bit US-ASCII Charset.In the simple application of data security processing in enterprise software development, using base64 encoding feature to make simple data encryption and decryption application. Its characteristic is that the operation is relatively simple, less operation, high efficient. Analysis from the strict definition, base64 is not a complete sense of data encryption, because the conversion process does not generate a key, can only conversion and recombination of the code. The base64 encoding standard is A – Z, a – z, 0 – 9, +, and /, as long as the program dynamically changes the order. The same characters will be encoded differently depending on their position, and the probability of being cracked will be greatly reduced to improve data security.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104077 |

G. Kalaiarasi, M. Maheswar, Prathima Devadas, M. Selvi[5]present a methodology for identifying near-duplicate images from the image set for an inquiry image by the user from the image features are extracted initially using a neural network called pulse coupled neural network similarity measures are calculated once the features are extracted. In this Hellinger, the coefficient is computed the near duplicate are recognized with help of a relation between the inquiry image and the image in the image set to reduce redundancy it is vital of the image. When compared to any other traditional system the accuracy and detection are enhanced in the PCNN-NDD system.The detection of Near Duplicate Images is completed by feature extraction utilizing Pulse Coupled Neural Network and computation of similarity utilizing Hellinger Correlation Coefficient. The outcomes show that the images are recognized appropriately.

Saehoon Kim [6] showed a scalable solution to discover near duplication on millions of images the method generates cluster seeds initially with min hashing for dividing and concurring the images then removable positive images are carried out by growing the cluster seeds value by making use of design growing function vitally. The basic components of seed growing are used to generate a different signature for candidate image removable which has only one common visual word with a seed cluster that is bottom -k min hash with help of the method precision and near-duplicate cluster can be discovered.

Bin Wang, Zhiwei Li [7] An replicate image detection algorithm for retrieving visual images in a set of images has been discussed. First, the k-bit hash code for each image is calculated i.e. each image is converted to a k-bit hash code according to its content and then makes duplicate image detection with hash codes only. In this, we have a tendency to propose an associate degree algorithmic rule with efficiency and effectively notice visually duplicate pictures in an exceedingly giant set of images. we have a tendency to 1st calculate a K-bit (K32) hash code for each image and conduct the duplicate image detection with only the hash codes. as a result of the hash codes area unit's terribly compact illustration of the image content, the detection method is incredibly quick. The experiments show the planned algorithmic rule will notice duplicate pictures with high exactitude and therefore the time price for grouping a hundred,000 pictures is a smaller amount than zero.4 second. Since we are able to represent pictures in such a compact type, one future work is to cluster the similar pictures to additional present additional organized results to users.

Maneesha,Indraveer Chana[8] presents method Brisky, Locality sensitive hashing.For features and extraction, BRSK and LSH are used for targeting purposes that map out the same image in the same bucket and use the hamming distance to view the same images. The drawback of using this is it is less accurate.

Nidhi B Sankhe, Shweta Singh, Sumedh Pundkar[9] propose a project that deletes duplicate media the techniques involved and used in this where SHA(secure hash algorithm) and dhash(difference hash algorithm). Deduplication improves storage utilization with higher reliability. Using SHA and Dhash the replicate image is deleted. The System helps in easy maintenance of data so that no duplicate files are saved.There is a rapid increase in the size of data which lead to heat energy consumption, duplicate data, and inconsistent data organization. In this paper, we propose a framework in order to fulfill a balance between changing storing efficiency and performance improvement in the system. The proposed system is capable of handling scalability problems by removing duplicate data. Deduplication aids in saving storage space. This project helps in easy maintenance of data so that no duplicate files are saved. It works for text, images, audio and video. With the evolution, storage resources of commodity machines can be efficiently utilized.

N. Jayanthi, S. Indu[10]Based on the feature detection and extraction techniques mentioned higher, we've got seen thatSURF algorithmic rule is the one of the simplest various for image matching issues. during this report, wehave mentioned varied vital techniques like Blob detection algorithmic rule, example matchingmethod, SIFT, and SURF algorithmic rules. we've got seen that the Blob detection algorithmic rule limits the U.S. to thenumber of gestures with lower accuracy and slower output. example matching techniques are slightly higher with intermediate quality and accuracy. but with an increasing range oftemplates, the potency and also the output of the algorithmic rule are adversely affected. In support of thestatements, we've got placed the tested pictures and also the corresponding outputs. Finally, the SURF algorithmis mentioned with its intermediate steps alongside the tested pictures and their outputs. We havetested the algorithmic rule available gestures, distinguishing objects out of an image, letters, and words fromtexts in English and Tamil language and located that it worked with nice accuracy and quicker speed.We additionally tested it by rotating and scaling the objects and located out that the algorithmic rule showed correctresults in the ninetieth of the cases. SURF algorithmic rule has pointed out to be one among the foremost strong featuredetection technique but it's sure limitations too like just in case of terribly low lit. images, distinctive the objects would be a bit tough, which is able to be Associate in Nursing improvementdirection for the longer-term work.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104077 |

SR.NO	AUTHORS	TECHNIQUES INVOLVED	ADVANTAGES	DRAWBACKS
1	Y. Ke, Rahul Suktankar	steady detection points (Dog Detection), local descriptors (PCAsift), and restrained data (LSH).	efficient data layout for interactive operations	using a part-based approach has to query hundred to thousand of features at an instance which could make it slow.
2	K.K.Thyagharajan G.Kalaiarasi	Pixel-based feature Extraction Method	of less relevancy to the user	1.Complexity in NDD 2.Computational complexity 3.Optimum values
3	Ammar Asaad, Ali Adil ,Yassin Alamri	Huffman code and hash function MD5	generates unique code for each image	Low storage space is the most important issue
4	Somchai, Wen Dong	Base64 Encoding	operation is relatively simple, less operation, high efficient	the probability of being cracked will be greatly reduced to improve data security.
5	M. Maheswar, Prathima Devadas, M. Selvi	pulse coupled neural network similarity measures	the accuracy and detection are enhanced	the image in the image set to reduce redundancy it is vital of the image.
6	Saehoon Kim	min hashing for dividing and concurring the images	discover near duplication on millions of images	removable positive images are carried out by growing the cluster seeds value
7	Bin Wang, Zhiwei Li	k-bit hash code for each image	the detection method is incredibly quick.	the hash codes area unit's terribly compact illustration of the image content
8	Maneesha,Indraveer Chana	Brisky, Locality sensitive hashing	Less time consuming	it is less accurate.
9	Nidhi B Sankhe, Shweta Singh, Sumedh Pundkar	Using SHA and Dhash	Deduplication improves storage utilization with higher reliability	A rapid increase in the size of data which lead to heat energy consumption, duplicate data, and inconsistent data organization.
10	N. Jayanthi, S. Indu	SURF algorithmic rule	it worked with nice accuracy and quicker speed	slightly higher with intermediate quality and accuracy

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 4, April 2022||

| DOI:10.15680/IJIRSET.2022.1104077 |

III. CONCLUSION

Huffman code to generate unique code for each image and save this code in image in the index file and use it to rid of a duplicate from the device's storage. Removing images that have the same code and maintaining one copy. The proposed scheme gets a good result in removingthese duplicate images fast. We applied it to two devices (Galaxy A5 [2017] SM-A520F, HUAWEI CUN-U29). As a final result, the performance of the device's CPU is better thanmaking the device work in the best way. In future work, in relation to the image matching schemes, the user may add something to the image, and as a result, a new image is obtained, but it is still the same image with a simple change.This paper includes a surveyof replicate detection and deletion of media. Various approaches and techniques involved to find the same images are studied. Some of these techniques have their own limitations and drawbacks. Also, the survey concentrates on base64 and dhash techniquesfor the detection of replicate images,pdf, and audio files. Several systems have been implemented with different approaches to detect those duplicate media. This paper presents a review of researchon duplication detection using several methods and techniques. The related survey provides guidance to our work.

REFERENCES

1. Yan Ke, Rahul Sukthankar, Larry Huston "Efficient near-duplicate Detection and Sub-image Retrieval Proceedings of ACM International Conference on Multimedia (MM)", 2004
2. K.K.Thyagharajan,G.Kalaiarasi " A Review on Near-Duplicate Detection Of Images using Computer Vision Techniques '', 02 January 2020.

3 Ammar Asaad ,Ali adil Yassin " A New Scheme For Removing Duplicate Files Frome Smart Mobile Devices", August 2019 4Somchai, Wen Dong, "Research on Base64 Encoding Algorithm and PHP Implementation", International Conference on Geoinformatics, 06 December 2018.

5 G. Kalaiarasi, M. Maheswar, Prathima Devadas Amruta Pise, S.D.Ruikar, "Text Detection and Recognition in Natural Scene Images", IEEE International Conference on Communication and Signal Processing, 2014.

6Saehoon Kim, Xin-Jing Wang, Lei Zhang, and Seungjin Choi. 2015. Near duplicate image discovery on one billion images. In Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. IEEE, 943–950

7 Bin Wang, Zhiwei Li "LARGE-SCALE DUPLICATE DETECTION FOR WEB IMAGE SEARCH"in Proc.IEEE Int.Conf.Multimedia Expo,July 2006

8. Maneesha, Inderveer Chana, "Searching of Duplicate Images in Cloud Database", International Conference on Inventive Computation Technology (ICICT), 19 January 2017.

9. Nidhi B Sankhe, Shweta Singh, Sumedh Pundkar "Using SHA and Dhash the replicate image is deleted",07 July 2021.

10. N. Jayanthi, S. Indu **"**Comparison of Image Matching Techniques",October 2018.

11 T. Kalaiselvi, K.Rahimunnisa, "Pulse Coupled Neural Network-based Duplicate Detection ", Detection of Image, 03 November 2018.

12 V. B. Nemirovsiy, A.K. Stoyanov, "Duplicate Image Recognition", IEEE, 2014.

13 G.Kalaiarsai, "Detecting Duplicate Image Using Hellinger Correlation Coeefifcient", Internal Journal of Advanced Research Engineering, 06 June 2020.

14 Sreedhar Kumar, Gunashree, "A New approach of Multilevel Unsupersived Clustering for Detecting Replication Level in Imageset", 2020.

15.https://www.base64decode.org/

16 .https://base64.guru/learn/base64-algorithm

17 .http://www.hackerfactor.com/blog/?/archives/529-Kind-of-Like-That.html

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details