

Project Report

On

**DELETION OF TWO FOLD DATA USING IMAGE
PROCESSING**

Submitted for partial fulfillment of requirement for the degree of

BACHELOR OF ENGINEERING

(Computer Science and Engineering)

Submitted By

Anuja Raghuwanshi

Likhita Vyas

Vaishnavi Pund

Under the Guidance of

Dr. G. R. Bamnote



**Department of Computer Science & Engineering,
PRM Institute of Technology & Research, Badnera.**

2019-2020



Department of Computer Science & Engineering

Prof. Ram Meghe Institute of Technology & Research,
Badnera 2019 - 2020

CERTIFICATE

This is to certify that the Project (8KS07) entitled

DELETION OF TWO FOLD DATA USING IMAGE PROCESSING

is a bonafide work and it is submitted to the

Sant Gadge Baba Amravati University, Amravati By

Anuja Raghuwanshi

Likhita Vyas

Vaishnavi Pund

*in the partial fulfillment of the requirement for the degree of Bachelor of
Engineering in Computer Science & Engineering , during the academic
year 2019-2020 under my guidance .*

Dr. G. R. Bamnote

Guide

*Department of Computer Sci. & Engg.
PRM Institute Of Technology & Research,
Badnera*

Dr. G. R. Bamnote

Head,

*Department of Computer Sci. & Engg.
PRM Institute Of Technology & Research,
Badnera*

External Examiner

ACKNOWLEDGEMENT

With great pleasure we hereby acknowledge the help given to us by various individuals throughout the project. This Project itself is an acknowledgement to the inspiration, drive and technical assistance contributed by many individuals. This project would have never seen the light of this day without the help and guidance we have received.

We would like to express our profound thanks to Dr. G. R. Bamnote for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project. We would also thank the faculties of the Department of Computer Science & Engineering, for their kind co-operation and encouragement which help us in completion of this project. We owe an incalculable debt to all staffs of the Department of Computer Science & Engineering for their direct and indirect help.

Our thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities.

We extend our heartfelt thanks to our parents, friends and well wishers for their support and timely help. Last but not the least; we thank the God Almighty for guiding us in every step of the way.

Anuja Raghuwanshi _____

Likhita Vyas _____

Vaishnavi Pund _____

TABLE OF CONTENTS

	LIST OF FIGURES	[3]
	LIST OF SCREENSHOTS	[4]
	ABSTRACT.....	[5]
1	INTRODUCTION	1
	1.1 Overview	1
	1.2 Problem Statement	3
	1.3 Project Objective	4
	1.4 Need of Application	6
2	LITERATURE REVIEW	7
	2.1 Review.....	7
	2.2 Related Work.....	8
	2.3 Image Processing Algorithms and Deduplication.....	9
	2.4 Summary and Description	11
3	SYSTEM ANALYSIS	13
	3.1 Problem Statement	13
	3.2 System Requirements	13
	3.2.1 Eclipse Software	13
	3.2.2 Apache Tomcat Server	14
	3.2.3 MySQL Database.....	15
	3.3 Technologies Involved	16
	3.3.1 Jakarata Server Pages(JSP).....	16
	3.3.2 Bootstarp	18
	3.3.3 JavaScript.....	19
4	SYSTEM ARCHITECTURE	20
	4.1 Image Processing.....	20
	4.2 Data Flow Diagram	21
	4.2.1 Image Compression	22
	4.2.2 Image Segmentation.....	23
	4.2.3 Image Hashing and Calculating Bit	24

5	IMPLEMENTATION AND RESULT	26
5.1	Implementation.....	26
5.1.1	Database Connectivity to Java Application	26
5.1.2	Web Scrapping	26
5.1.3	Image Comparison	27
5.1.4	K-Means Algorithm for Segmentation	27
5.2	Application of Image Processing.....	29
5.3	Result	31
5.3.1	Checking for Duplicate Image.....	31
5.3.2	Checking for Duplicate Texts	36
6	CONCLUSION AND FUTURE SCOPE.....	38
6.1	Conclusion.....	38
6.2	Future Scope.....	39
6.2.1	Future Scope in Data Deuplication	39
6.2.2	Future in Image Processing.....	39
	REFERENCES	41

LIST OF FIGURES

Figure 1.1: Object Detection	3
Figure 1.2: Two Fold Data with Image Processing	5
Figure 2.1: Image Processing	7
Figure 3.1: Application Overview	15
Figure 3.2: JSP Model 2 Architecture	17
Figure 4.1: Digital Image Processing	20
Figure 4.2: Data Flow Diagram	21
Figure 4.3: Image Segmentation.....	23
Figure 4.4: RGB Value Calculation.....	24
Figure 4.5: Application Flowchart.....	25

LIST OF SCREENSHOTS

Screenshot 1: Login and Registration.....	31
Screenshot 2: Option for Uploading Image and Texts	31
Screenshot 3: Image Upload.....	32
Screenshot 4: Segmentation Process	32
Screenshot 5: Previously Calculated RGB Value of Images.....	33
Screenshot 6: Comparing Images in Database	33
Screenshot 7: Comparison Graph of RGB Values	34
Screenshot 8: Duplicate Copy Detected	34
Screenshot 9: Showing Result	35
Screenshot 10: Previously Uploaded Images with Time	35
Screenshot 11: Uploading Text	36
Screenshot 12: Result (No Duplicate Data).....	36
Screenshot 13: Previously Uploaded Texts with Time	36
Screenshot 14: Training Data Set	36

ABSTRACT

The amount of data we produce every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at our current pace, but that pace is only accelerating with the growth of the Internet of Things (IoT). Over the last two years alone 90 percent of the data in the world was generated. The main contribution is played by social media apps. In every minute millions of people are using social sites and apps for sharing different sorts of medias like photos videos even text messages which consume massive amount of storage. These media might have duplicate existence in this large cloud memory. The project aims to find out replica of these images present in user's database. In this project, image processing and concepts of machine learning can lead us to solve the problem of replication.

The project will generate the RGB value of uploaded image and compare it with the other images in database. The images which have nearly same value will be compared with the uploaded by using K-means clustering algorithm and mapping technique which will detect whether the image is present in database or not. The same procedure will be followed with text using SHA algorithms. This project will help us to use our storage efficiently without any wastage of memory.

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW:

Multimedia is increasingly becoming the “biggest big data” as the most important and valuable source for insights and information. An image is visual representation of information that cannot be represented by many words. Because of immense growth in technology many high end software available which can harm the sensitive information in images.

Recently, with the development of Internet and the popularity of digital products, the volume of global digital resource is growing at an alarming rate. For examples, in 2007, for the first time ever, the total volume of digital resource exceeded the global storage capacity. It is estimated that by 2011 only half of the digital information will be stored .Hence, it is impossible to solve the data explosion problem by blindly increasing storage devices. In order to solve the requirement of storage space, Kai Li Professor of Princeton University presented a new technology called global compression technology or De-duplication. De-duplication can identify redundant data, eliminate all but one copy, and create local pointers to the information that users can access.

Data storage costs can be significant depending on the type of data you store, and duplicate records can occupy a lot of storage space. While you wouldn't expect a CRM record to be duplicated hundreds or even thousands of times, consider other types of data that a company might store.

Consider an email attachment. Let's say that a 1 MB email attachment was sent by 100 people within your company. With 100 instances of the attachment saved in your database, this requires 100 MB of storage space in total. By removing duplicate entries through data de-duplication, only one instance of the attachment is required to be stored, resulting in a total of 1 MB of space being occupied.

This goes for any user if the images are duplicated it will cost upto 80% of the usage space which affects the performance of device.

Nowadays people are worried about the storage space and the great amount of the memory space is occupied by the media. The moreover it exceeds the storage space if the redundant images are being stored. To improve the amount of storage space we have developed an application for the user generated platform called “Data De-Duplication Determination System.”

Data de-duplication refers to a series of techniques and strategies that are used to eliminate redundant data in a database. In a successful de-duplication campaign, extra duplicate copies of a record are deleted, leaving only one copy of the record in question. Data de-duplication systems analyze byte patterns to identify duplicate copies of the same record. Typically, the extra copies are merged or deleted and replaced with a reference that routes back to the remaining permanent record.

In the existing system the storage is increasing due to duplication of images. Huge amount of data is being created and absorbing the storage space available on cloud. This project will detect the duplicate media (images, text messages) and remove it from user's database. Firstly, the program will calculate the RGB (Red Green Blue) of uploaded image. Then Images which are already present and have nearly same value will be compared with the uploaded by using K-means clustering algorithm and mapping technique which will detect whether the image is present in database or not. The same procedure will be followed with text using SHA(Secure Hash Algorithm).

We can divide or partition the image into various parts called segments. It's not a great idea to process the entire image at the same time as there will be regions in the image which do not contain any information. By dividing the image into segments, we can make use of the important segments for processing the image. That, in a nutshell, is how image segmentation works.

Object detection builds a bounding box corresponding to each class in the image. But it tells us nothing about the shape of the object. We only get the set of bounding box coordinates. We want to get more information this is too vague for our purposes. Image segmentation creates a pixel-wise mask for each object in the image. This technique gives us a far more granular understanding of the object in the image. An image is a collection or set of different pixels. We group together the pixels that have similar attributes using image segmentation.

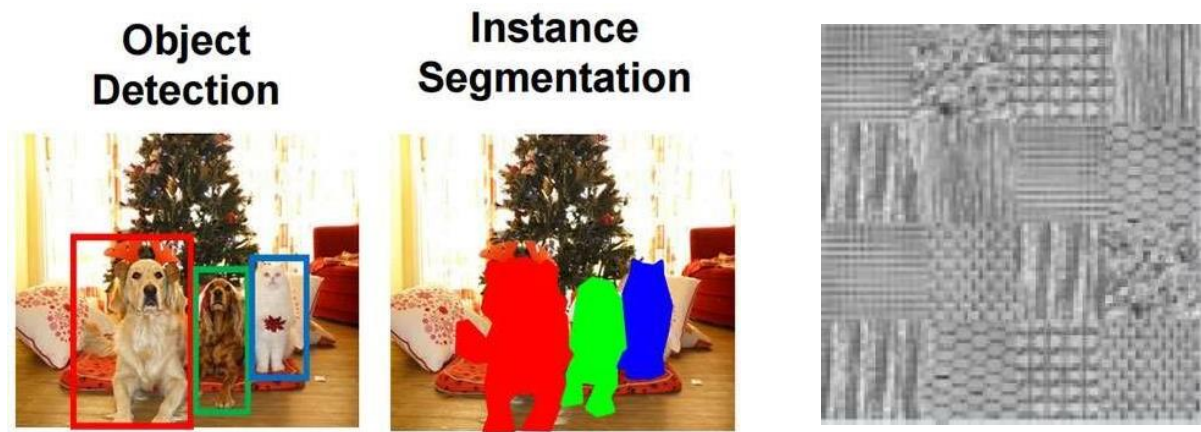


Figure 1.1 Object Detection

1.2 PROBLEM STATEMENT:

Several applications are accepting the same image shared by multiple users which are consuming the memory space. These redundant images are creating some part of issue in amount of the storage space required to a particular user. Due to large space required by the media if there are duplicate images the memory space consumed is very large and user is unable to store data efficiently. The impact of storage is caused directly on the device as it stops working effortlessly and causes buffering while accessing the device.

1.3 PROJECT OBJECTIVES:

We are implementing technique for eliminating redundant data in a data set in such a way that the duplicate images in set will be deleted before the memory space is occupied by that image this will result in no consumption of memory space for duplicate images and the user can conveniently manage the space and avoid the bit amount of storage problems.

(i) In the process of de-duplication, the problem is solved as same images of the same data are deleted, leaving only one image to be stored.

(ii) The purpose of this project is to research current image processing tools and create a simple, easy image extracting algorithms and flexible image processing widget based on the Java Advanced Imaging (JAI) API.

(iii) The project will develop the set of activities on Twofold media and image processing.

(a) Twofold Media:

Being used in the scientific terms, twofold data is duplicate data or redundant data where data can be in the form of media like images or text. Redundancy means having multiple copies of same data in the dataset. Storage-based data deduplication reduces the amount of storage needed for a given set of files. It is most effective in applications where many copies of very similar or even identical data are stored on a single disk—a surprisingly common scenario. In the case of data backups, which routinely are performed to protect against data loss, most data in a given backup remain unchanged from the previous backup.

(b) Image Processing:

Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Image processing basically includes the following three steps:

(1) Importing the image via image acquisition tools:

Importing the images in the primary dataset by downloading or moving the images. After moving the images in the primary dataset we will be trying to import duplicate copy of the image through the process created by using algorithms.

(2) Analysing and manipulating the image

The process will detect the duplicate image and will be analyzed by clustering, segmentation and CLD formula using mapping techniques and checking the rate of duplicacy and calculating the RGB value. One the process is completed we get the message whether the image is moved in the primary dataset or deleted due to the redundancy.

(3) Output in which result can be altered image or report that is based on image analysis

Duplicate image detected will not be moved in the primary dataset and it will be shown in the other duplicate set of images for the rechecking that which image is deleted. And if the image is not matched with the images in the primary dataset it will be moved to the dataset. As duplicate image is found the folder is empty.

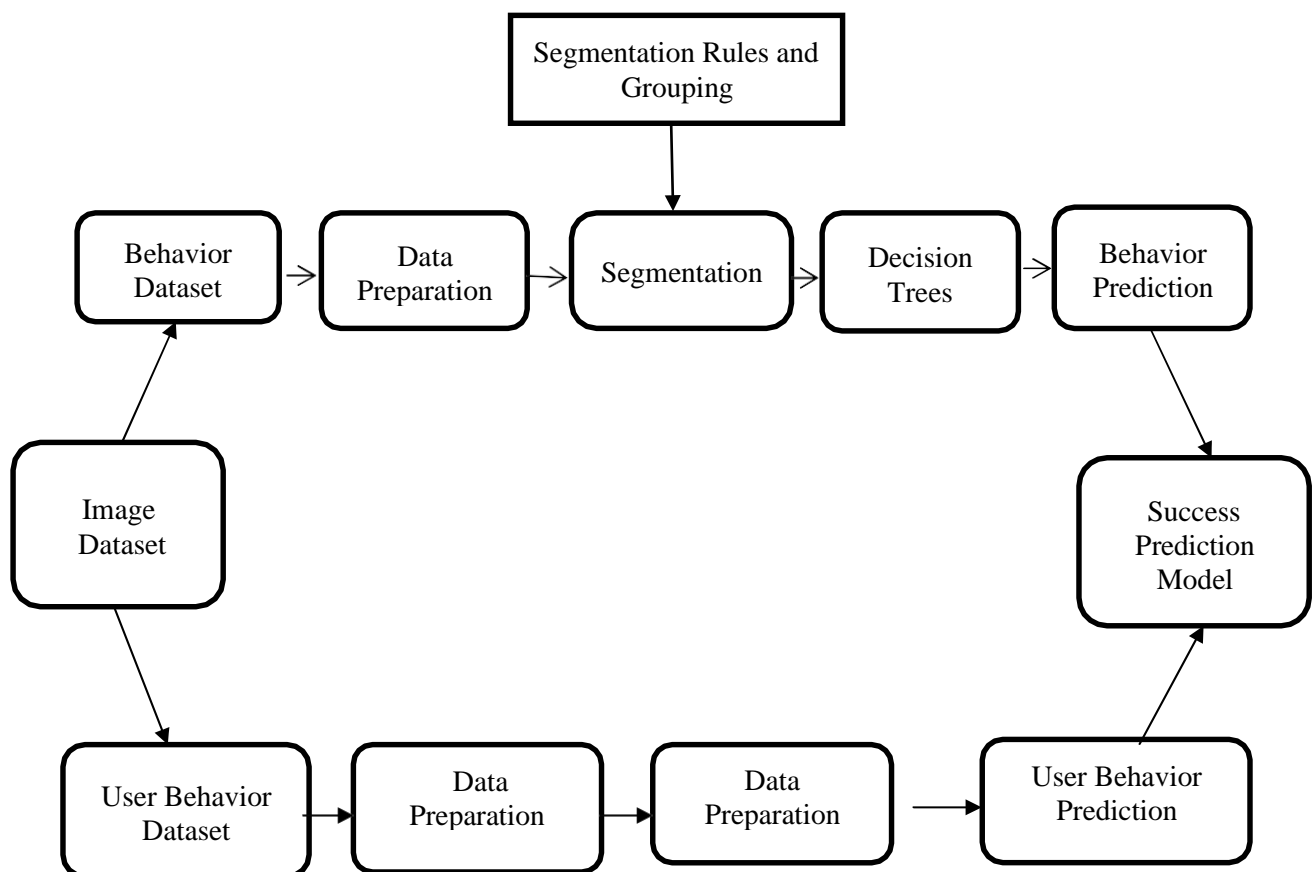


Figure 1.2 Twofold Data with Image Processing

The process is flow is as shown in the above diagram i.e. how the image is processed and the message is retrived by the data duplication system. When the redundant image will be detected the process will be followed by the two datasets user data which will be training images and the device data i.e. uploaded images.

1.4 NEED OF APPLICATION:

Our application will help the user to save their memory space and access the data easily without rendering the device and smooth access. As the data is not directly stored in the memory it directly access with uploaded source it makes our application well featured. The Web Application is user friendly with its easy understandable user interface.

CHAPTER 2

LITERATURE SURVEY

2.1 REVIEW:

Digital Image Processing (DIP), more commonly known as Image Processing deals with manipulation of digital images using a digital computer. It is a sub field of signals and systems but focuses particularly on images. DIP focuses on developing a computer system that is able to perform processing on an image. The input of such system is a digital image. The system processes the image using efficient algorithms, and gives an image as an output.

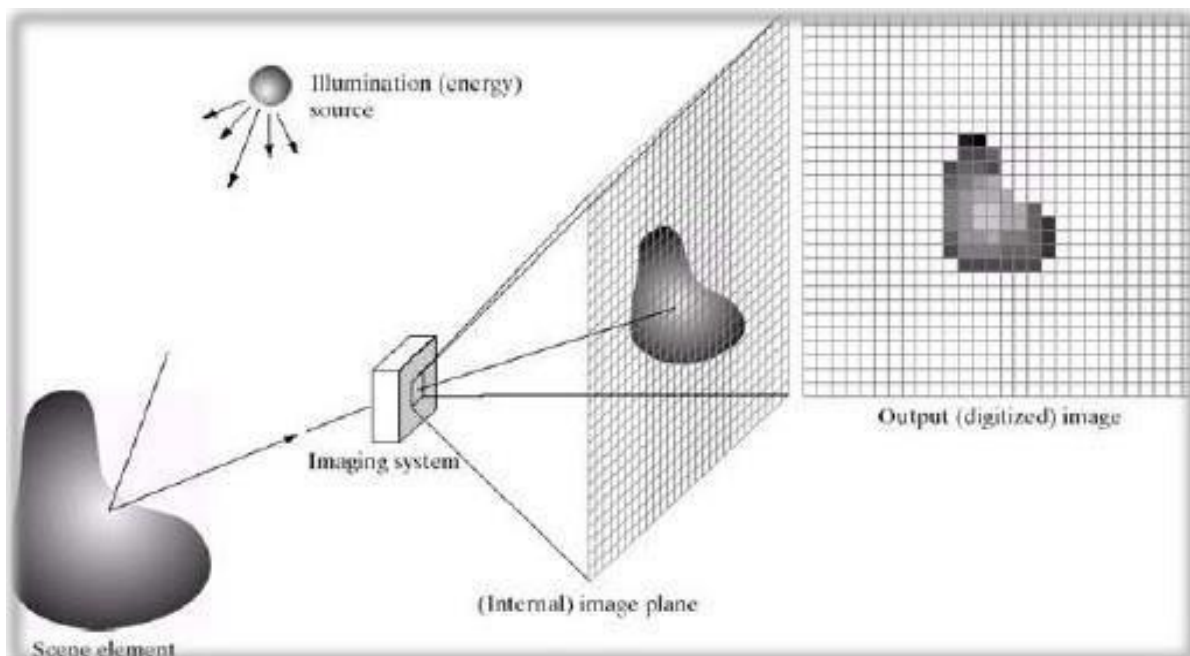


Figure 2.1 Image Processing

An image is nothing more than a two dimensional signal. It is defined by the mathematical function $f(x,y)$ where x and y are the two co-ordinates horizontally and vertically. The value of $f(x,y)$ at any point gives the pixel value at that point of an image.

Many of the techniques of digital image processing, or digital picture processing as it often was called, were developed in the 1960s, at Bell Laboratories, the Jet Propulsion Laboratory, Massachusetts Institute of Technology, University of Maryland, and a few other research facilities, with application to satellite imagery, wire-photo standards conversion, medical imaging, videophone, character recognition, and photograph enhancement.

The purpose of early image processing was to improve the quality of the image. It was aimed for human beings to improve the visual effect of people. In image processing, the input is a low-quality image, and the output is an image with improved quality. Common image processing include image enhancement, restoration, encoding, and compression. The first successful application was the American Jet Propulsion Laboratory (JPL). They used image processing techniques such as geometric correction, gradation transformation, noise removal, etc. on the thousands of lunar photos sent back by the Space Detector Ranger 7 in 1964.

2.2 RELATED WORK:

There have been many works done in the area of image segmentation by using different methods. And many are done based on different application of image segmentation. *K*-means algorithm is the one of the simplest clustering algorithm and there are many methods implemented so far with different method to initialize the center . And many researchers are also trying to produce new methods which are more efficient than the existing methods, and shows better segmented result. Some of the existing recent works are discussed here.

Pallavi Purohit and Ritesh Joshi⁴ introduced a new efficient approach towards *K*-means clustering algorithm. They proposed a new method for generating the cluster center by reducing the mean square error of the final cluster without large increment in the execution time. It reduced the means square error without sacrificing the execution time. Many comparisons have been done and it can conclude that accuracy is more for dense dataset rather than sparse dataset.

Alan Jose, S. Ravi and M. Sambath⁵ proposed Brain Tumor Segmentation using *K*-means Clustering and Fuzzy. *C*-means Algorithm and its area calculation. In the paper, they divide the process into three parts, pre-processing of the image, advanced *k*-means and fuzzy *c*-means and lastly the feature extraction. First pre-processing is implemented by using the filter where it improves the quality of the image. Then the proposed advance *K*-means algorithm is used, followed by Fuzzy *c*-means to cluster the image. Then the resulted segment image is used for the feature extraction for the region of interest. They used MRI image for the analysis and calculate the size of the extracted tumor region in the image.

Madhu Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa⁶ proposed Enhancing K -means clustering algorithm with improved initial center. A new method for finding the initial centroid is introduced and it provides an effective way of assigning the data points to suitable clusters with reduced time complexity. They proved their proposed algorithm has more accuracy with less computational time comparatively original k -means clustering algorithm. This algorithm does not require any additional input like threshold value. But this algorithm still initializes the number of cluster k and suggested determination of value of k as one of the future work.

K. A. Abdul Nazeer, M. P. Sebastian⁷ proposed an enhanced algorithm to improve the accuracy and efficiency of the k -means clustering algorithm. They present an enhanced k -means algorithm which combines a systematic method consisting two approaches. First one is finding the initial centroid and another is assigning the data point to the clusters. They have taken different initial centroid and tested execution time and accuracy. From the result it can be conclude that the proposed algorithm reduced the time complexity without sacrificing the accuracy of clusters.

2.3 IMAGE PROCESSING ALGORITHMS AND DEDUPLICATION:

In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

Image segmentation plays a significant role in computer vision. It aims at extracting meaningful objects lying in the image. Generally there is no unique method or approach for image segmentation. Clustering is a powerful technique that has been reached in image segmentation. Clustering is the process of making a group of abstract objects into classes of similar objects.

There are various machine learning algorithms that are used in the image processing realm. However, in the Project we have mainly used the following two algorithms. They both different purposes and in their own sense are very useful and the backbone of the project.

SHA Algorithm:

Secure Hash Algorithms, also known as SHA, are a family of cryptographic functions designed to keep data secured. It works by transforming the data using a hash function: an algorithm that consists of bitwise operations, modular additions, and compression functions. The hash function then produces a fixed-size string that looks nothing like the original. These algorithms are designed to be one-way functions, meaning that once they're transformed into their respective hash values, it's virtually impossible to transform them back into the original data. A few algorithms of interest are SHA-1, SHA-2, and SHA-3, each of which was successively designed with increasingly stronger encryption in response to hacker attacks. SHA-0, for instance, is now obsolete due to the widely exposed vulnerabilities.

The SHA function is an algorithm that hashes data such as a text file into a fixed length variable known as a hash. This computed hash value is then used to verify copies of the original data.

A common application of SHA is to encrypting passwords, as the server side only needs to keep track of a specific user's hash value, rather than the actual password. This is helpful in case an attacker hacks the database, as they will only find the hashed functions and not the actual passwords, so if they were to input the hashed value as a password, the hash function will convert it into another string and subsequently deny access. Additionally, SHAs exhibit the avalanche effect, where the modification of very few letters being encrypted causes a big change in output; or conversely, drastically different strings produce similar hash values. This effect causes hash values to not give any information regarding the input string, such as its original length. In addition, SHAs are also used to detect the tampering of data by attackers, where if a text file is slightly changed and barely noticeable, the modified file's hash value will be different than the original file's hash value, and the tampering will be rather noticeable.

Hashing provides a more reliable and flexible method of data retrieval than any other data structure. It is faster than searching arrays and lists. In the same space it can retrieve in 1.5 probes anything stored in a tree that will otherwise take $\log n$ probes.

Unlike any other data structure hashing does not determine what speed you will get; with hashing one can choose the time-space tradeoff to make. You control either the speed by

picking the amount of space for the hash table; or you control the space by picking a speed of retrieval.

However hashing does not do a good job of sorting data, since it work by randomizing not ordering data. This is a disadvantage if you want to present an ordered list of the data, not just retrieve it.

2.4 SUMMARY AND DESCRIPTION:

Initially the emphasis of researchers was on the reduction of data footprint by using disk rather than using tape for backups. From the start of year 2003 virtual tape libraries (VTL) were extensively used by the industry but the real advent of deduplication in fact, started in a 1970's when companies use to store large amount of customer contact information on tapes and had to eliminate duplicate entries. But, today deduplication is not only doing value addition in backup systems but additionally it is optimizing IOPs, SSD and DRAM efficiencies also.

The various implementation of deduplication algorithm show that when a virtual machine disk is created which certain X amount of data already resting in it. It would not require to rewrite those X amount of data to the disk. Thus saving significant IOPs resources. Deduplication also improves the efficiency of SSD and DRAM as it able to maintain same piece of data (in case it is duplicate) in minimum space. This is significant saving as DRAM and SSD are more expensive than HDD. Then disaster recovery system also get more optimize with the aid of deduplication process. However, all these processes need highly efficient key management and optimized proof of ownership work flows for them to be successful.

Image processing is the use of a digital computer to process digital images through an algorithm. The generation and development of digital image processing are mainly affected by three factors: first, the development of computers; second, the development of mathematics (especially the creation and improvement of discrete mathematics theory); third, the demand for a wide range of applications in environment, agriculture, military, industry and medical science has increased.

The significance of clustering algorithms is to extract value from large quantities of structured and unstructured data. It allows you to segregate data based on their properties/features and group them into different clusters depending on their similarities.

Clustering algorithms have a variety of uses in different sectors. For example, you may need it for classifying diseases in medical science and classifying customers in the field of market research.

CHAPTER 3

SYSTEM ANALYSIS

3.1 PROBLEM STATEMENT:

Most enterprises are already painfully aware of the costs, errors and missed opportunities associated with duplicate data. Records for customers, suppliers, products, and more are duplicated in multiple systems due to a proliferation of operational systems and mergers and acquisitions. Often no mechanism exists to uniquely identify each entity across systems and no proactive steps are taken to prevent the creation of duplicate records. Our project deals with this massive problem on small sector of duplicates which caused by social media like whatsapp, facebook, etc. This project will help the user to manage their database and use it so effectively so that there will not be any duplicate media like documents, images and messages on their database.

3.2 SYSTEM REQUIREMENTS:

Our project has some software requirements as follows:

3.2.1 Eclipse Software:

Eclipse is an integrated development environment (IDE) used in computer programming. It contains a base workspace and an extensible plug-in system for customizing the environment. We used this software to develop our project because of it's compatibility with java programming language. The Eclipse SDK includes the Eclipse Java development tools (JDT), offering an IDE with a built-in Java incremental compiler and a full model of the Java source files. This allows for advanced refactoring techniques and code analysis. The IDE also makes use of a workspace, in this case a set of metadata over a flat file space allowing external file modifications as long as the corresponding workspace resource is refreshed afterward.

Eclipse implements the graphical control elements of the Java toolkit called Standard Widget Toolkit (SWT), whereas most Java applications use the Java standard Abstract Window Toolkit (AWT) or Swing. Eclipse's user interface also uses an intermediate graphical user interface layer called JFace, which simplifies the construction of applications based on SWT.

3.2.2 Apache Tomcat Server:

Apache Tomcat (called "Tomcat" for short) is an open-source implementation of the Java Servlet, JavaServer Pages, Java Expression Language and WebSocket technologies. Tomcat provides a "pure Java" HTTP web server environment in which Java code can run. We have used Tomcat 4.x which was released with Catalina (a servlet container), Coyote (an HTTP connector) and Jasper (a JSP engine). As our project is a web application for detecting duplicate images in user's own created database so we used this because it has also added user— as well as system-based web applications enhancement to add support for deployment across the variety of environments. It also tries to manage sessions as well as applications across the network.

Tomcat is building additional components. A number of additional components may be used with Apache Tomcat. These components may be built by users should they need them or they can be downloaded from one of the mirrors. We have also used its high-availability feature facilitate the scheduling of system upgrades (e.g. new releases, change requests) without affecting the live environment. This is done by dispatching live traffic requests to a temporary server on a different port while the main server is upgraded on the main port. It is very useful in handling user requests on high-traffic web applications.

3.2.3 MySQL Database:

MySQL is an open-source relational database management system (RDBMS). A relational database organizes data into one or more data tables in which data types may be related to each other; these relations help structure the data. SQL is a language programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.

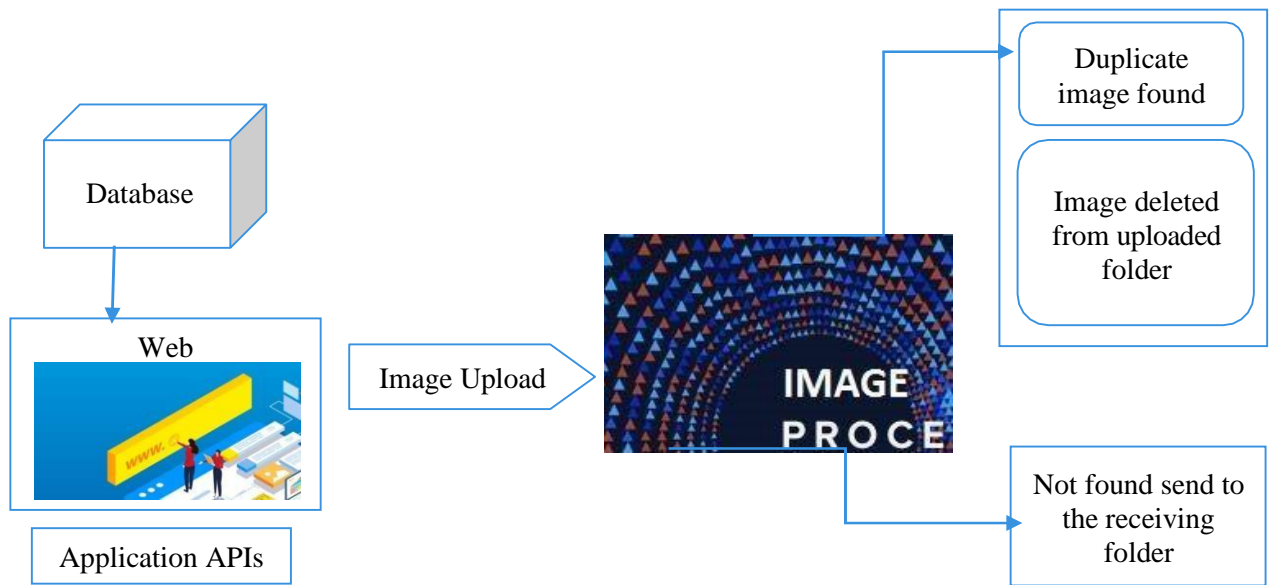


Figure 3.1 Application Overview

3.3 TECHNOLOGIES INVOLVED:

3.3.1 Jakarta Server Pages (JSP):

Jakarta Server pages is one of the original java web technology which is being widely used to create dynamic web pages that can connect to java backend. It is built on top of the Java Servlet specification. JSP may be viewed as a high-level abstraction of Java servlets. JSPs are translated into servlets at runtime, therefore JSP is a Servlet; each JSP servlet is cached and re-used until the original JSP is modified. Jakarta Server Pages can be used independently or as the view component of a server-side model–view–controller design, normally with JavaBeans as the model and Java servlets (or a framework such as Apache Struts) as the controller.

JSP allows Java code and certain predefined actions to be interleaved with static web markup content, such as HTML. The resulting page is compiled and executed on the server to deliver a document. The compiled pages, as well as any dependent Java libraries, contain Java bytecode rather than machine code. Like any other .jar or Java program, code must be executed within a Java virtual machine (JVM) that interacts with the server's host operating system to provide an abstract, platform-neutral environment.

JSPs are usually used to deliver HTML and XML documents, but through the use of OutputStream, they can deliver other types of data as well. The Web container creates JSP implicit objects like request, response, session, application, config, page, pageContext, out and exception. JSP Engine creates these objects during translation phase.

Architecturally, JSP may be viewed as a high-level abstraction of Java servlets. JSPs are translated into servlets at runtime, therefore JSP is a Servlet; each JSP servlet is cached and re-used until the original JSP is modified.

JSP can be used independently or as the view component of a server-side model–view–controller design, normally with JavaBeans as the model and Java servlets as the controller. The type of Model 2 architecture is shown in figure,

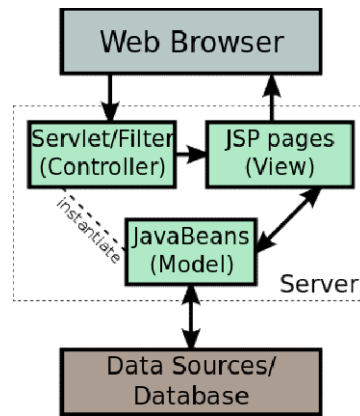


Figure 3.2 JSP Model 2 Architecture

JSP technology is the extension to Servlet technology. The main features of JSP technology are as follows:

- ❖ A language for developing JSP pages, which are text-based documents that describe how to process a request and construct a response
- ❖ An expression language for accessing server-side objects
- ❖ Mechanisms for defining extensions to the JSP language

Servlets provide URL mapping and request handling capabilities in your Java web applications. Request handling is the bread and butter of Java web application development. In order to respond to requests from the network, a Java web application must first determine what code will respond to the request URL, then marshal a response. Every technology stack has a way of accomplishing request-response handling. In Java, we use servlets (and the Java Servlet API) for this purpose. Think of a servlet as a tiny server whose job is to accept requests and issue responses.

3.3.2 Bootstrap:

Bootstrap is the most popular open-source framework full of useful and common classes to use in any project. It helps to develop responsive and mobile-first websites faster and easier. It is known for its faster and effortless responsive web development assistance, Bootstrap web design methodology utilize HTML and CMS based templates for user interface components like forms, navigations, alerts, buttons, typography in addition to optional JavaScript extensions.

Bootstrap is a web framework that focuses on simplifying the development of informative web pages (as opposed to web apps). The primary purpose of adding it to a web project is to apply Bootstrap's choices of color, size, font and layout to that project. As such, the primary factor is whether the developers in charge find those choices to their liking. Once added to a project, Bootstrap provides basic style definitions for all HTML elements. The result is a uniform appearance for prose, tables and form elements across web browsers. In addition, developers can take advantage of CSS classes defined in Bootstrap to further customize the appearance of their contents. For example, Bootstrap has provisioned for light- and dark-colored tables, page headings, more prominent pull quotes, and text with a highlight. Bootstrap also comes with several JavaScript components in the form of jQuery plugins. They provide additional user interface elements such as dialog boxes, tooltips, and carousels. Each Bootstrap component consists of an HTML structure, CSS declarations, and in some cases accompanying JavaScript code. They also extend the functionality of some existing interface elements, including for example an auto-complete function for input fields.

The most prominent components of Bootstrap are its layout components, as they affect an entire web page. The basic layout component is called "Container", as every other element in the page is placed in it. Developers can choose between a fixed-width container and a fluid-width container. While the latter always fills the width of the web page, the former uses one of the four predefined fixed widths, depending on the size of the screen showing the page:

Smaller than 576 pixels

576–768 pixels

768–992 pixels

992–1200 pixels Larger than

1200 pixels

Once a container is in place, other Bootstrap layout components implement a CSS Flexbox layout through defining rows and columns.

A precompiled version of Bootstrap is available in the form of one CSS file and three JavaScript files that can be readily added to any project. The raw form of Bootstrap, however, enables developers to implement further customization and size optimizations. This raw form is modular, meaning that the developer can remove unneeded components, apply a theme and modify the uncompiled Sass files.

3.3.3 JavaScript

JavaScript is one of the core technologies of the WWW (World Wide Web). It enables interactive web pages and is an essential part of web applications. It has application programming interfaces (APIs) for working with text, dates, regular expressions, standard data structures, and the Document Object Model (DOM). Almost all the websites and web browser uses JavaScript engines to execute client side page behavior. JavaScript engines were originally used only in web browsers, but they are now embedded in some servers, usually via Node.js. They are also embedded in a variety of applications created with frameworks such as Electron and Cordova.

CHAPTER 4

SYSTEM ARCHITECTURE

4.1 IMAGE PROCESSING:

Image processing is a method to convert an image into digital form and perform some operations on it, in order to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which input is image, like video frame or photograph and output may be image or characteristics associated with that image.

Image processing system includes treating images as two dimensional signals while applying already set signal processing methods to them. It is among rapidly growing technologies today, with its applications in various aspects of a business. Image Processing forms core research area within engineering and computer science disciplines too.

The simple definition of image processing refers to the processing of a digital image, i.e. eliminating the noise and kind of anomalies existing in an image using the digital computer. Image processing is way to perform some operations on an image to acquire an improved image or to cutting some useful information from it.

- Image processing basically includes the following three steps :
 1. Importing the image with optical scanner or by digital photography.
 2. Analyzing and manipulating the image which includes data compression and image enhancement and spotting patterns that are not to human eyes like satellite photographs.
 3. Output is the last stage in which result can be altered image or report that is based on image analysis.



Figure 4.1 Digital Image Processing

4.2 DATA FLOW DIAGRAM (DFD):

While using social connecting platforms to interact and connect with people, a huge amount of data is generated in the mean process. This data could be in any form like images, videos, text messages or any documents. These forms of media consumes massive amount of storage in user's phone memory or their database, which leads to buying costly cloud storage. The immense amount of storage is used by duplicate media present in our systems, this web application will help us to remove those duplicate images and files which are consuming the space in memory. The following figure shows procedure of deletion of duplicate images in user's database,

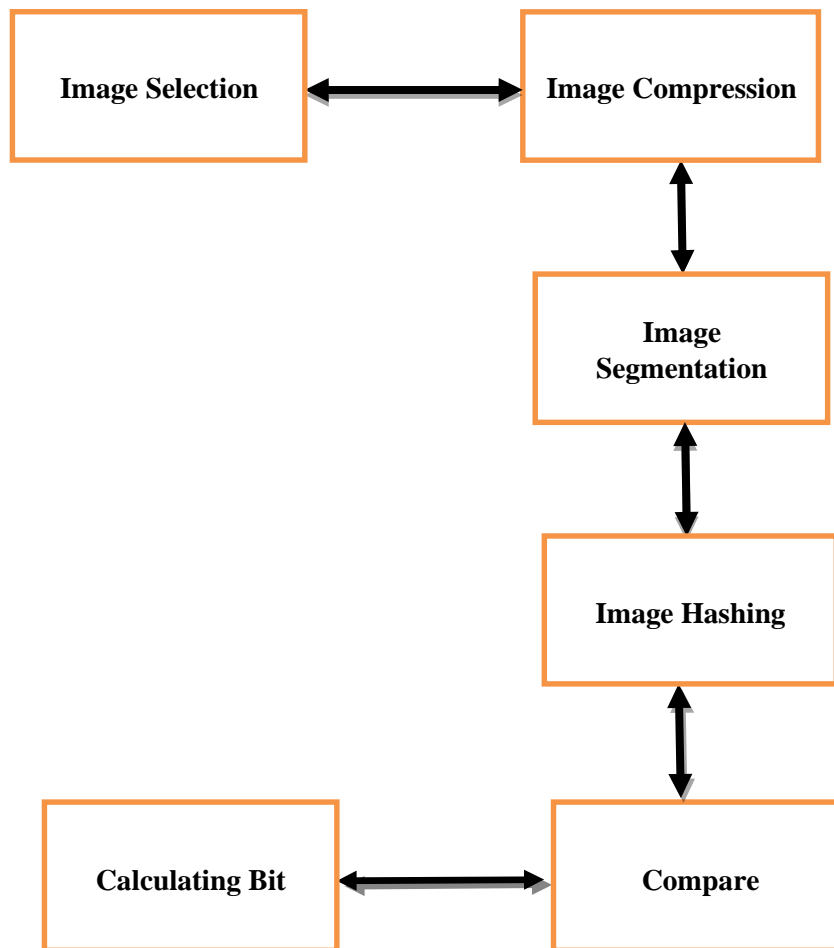


Figure 4.2 Data Flow Diagram

4.2.1 Image Compression:

Compression, as the name implies, deals with techniques for reducing the storage required to save an image, or the bandwidth required to transmit it. Although storage technology has improved significantly over the past decade, the same cannot be said for transmission capacity. This is true particularly in uses of the Internet, which are characterized by significant pictorial content. Image compression is familiar (perhaps inadvertently) to most users of computers in the form of image file extensions, such as the jpg file extension used in the JPEG (Joint Photographic Experts Group) image compression standard. It is a type of data compression applied to digital images, to reduce their cost for storage or transmission. We are using compression because we have to use the storage effectively. Basically, there are two types of compression.

- Lossy Image Compression
- Lossless Image Compression

Lossy compression or irreversible compression is the class of data encoding methods that uses inexact approximations and partial data discarding to represent the content. These techniques are used to reduce data size for storing, handling, and transmitting content. The different versions of the photo of the cat to the right show how higher degrees of approximation create coarser images as more details are removed. This is opposed to lossless data compression (reversible data compression) which does not degrade the data. The amount of data reduction possible using lossy compression is much higher than through lossless techniques.

Lossless compression is a class of data compression algorithms that allows the original data to be perfectly reconstructed from the compressed data. By contrast, lossy compression permits reconstruction only of an approximation of the original data, though usually with greatly improved compression rates (and therefore reduced media sizes). Lossless data compression is used in many applications. For example, it is used in the ZIP file format and in the GNU tool gzip. It is also often used as a component within lossy data compression technologies (e.g. lossless mid/side joint stereo preprocessing by MP3 encoders and other lossy audio encoders). Lossless compression is used in cases where it is important that the original and the decompressed data be identical, or where deviations from the original data would be unfavourable.

4.2.2 Image Segmentation:

Segmentation procedures partition an image into its constituent parts or objects. In general, autonomous segmentation is one of the most difficult tasks in digital image processing. A rugged segmentation procedure brings the process a long way toward successful solution of imaging problems that require objects to be identified individually. On the other hand, weak or erratic segmentation algorithms almost always guarantee eventual failure. In general, the more accurate the segmentation, the more likely recognition is to succeed.

Representation and description almost always follow the output of a segmentation stage, which usually is raw pixel data, constituting either the boundary of a region (i.e., the set of pixels separating one image region from another) or all the points in the region itself. In either case, converting the data to a form suitable for computer processing is necessary. The first decision that must be made is whether the data should be represented as a boundary or as a complete region. Boundary representation is appropriate when the focus is on external shape characteristics, such as corners and inflections. Regional representation is appropriate when the focus is on internal properties, such as texture or skeletal shape. In some applications, these representations complement each other. Choosing a representation is only part of the solution for transforming raw data into a form suitable for subsequent computer processing. A method must also be specified for describing the data so that features of interest are highlighted.

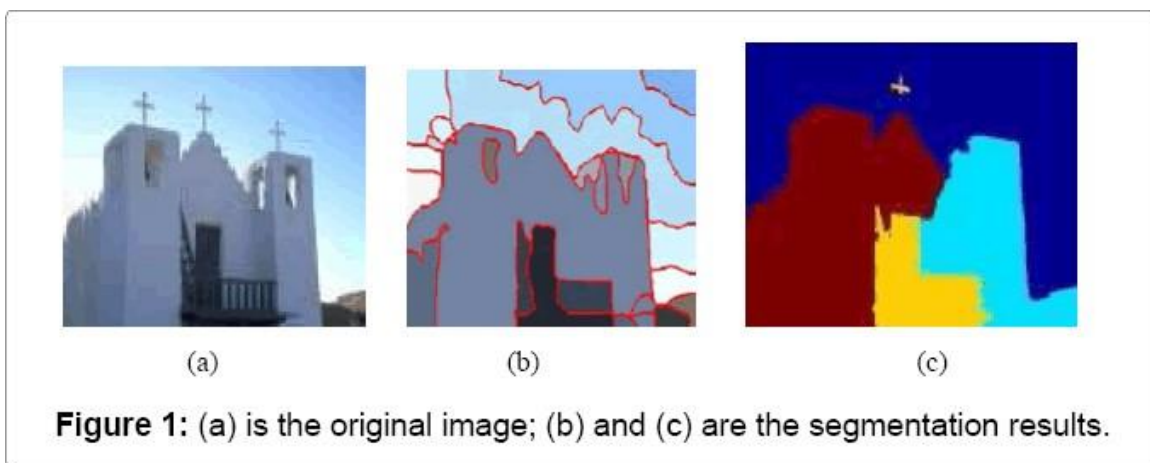


Figure 4.3 Image Segmentation

4.2.3 Image Hashing and Calculating Bit:

Image hashes tell whether two images look nearly identical. This is different from cryptographic hashing algorithms (like md5, sha-1) where tiny changes in the image give completely different hashes. In image fingerprinting, we actually want our similar inputs to have similar output hashes as well. The image hash algorithms (average, perception, difference, wavelet) analyse the image structure on luminance (without color information). The color hash algorithm analyses the color distribution and black & gray fractions (without position information).

In project, Image hashing is used for calculating the RGB (Red Green Blue) value of an uploaded image which will further used for comparison with the images present in database. Whenever the user will upload an image, Hashing algorithm will generate a unique average RGB value for each entry. This value will be compared with other and the image which have same RGB value to the uploaded image, that image will be deleted from database.

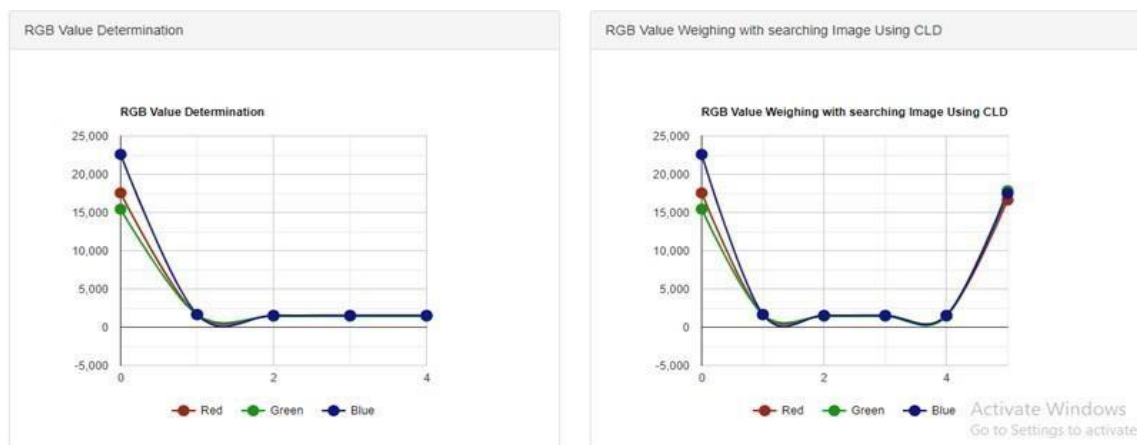


Figure 4.4 RGB Value Calculation

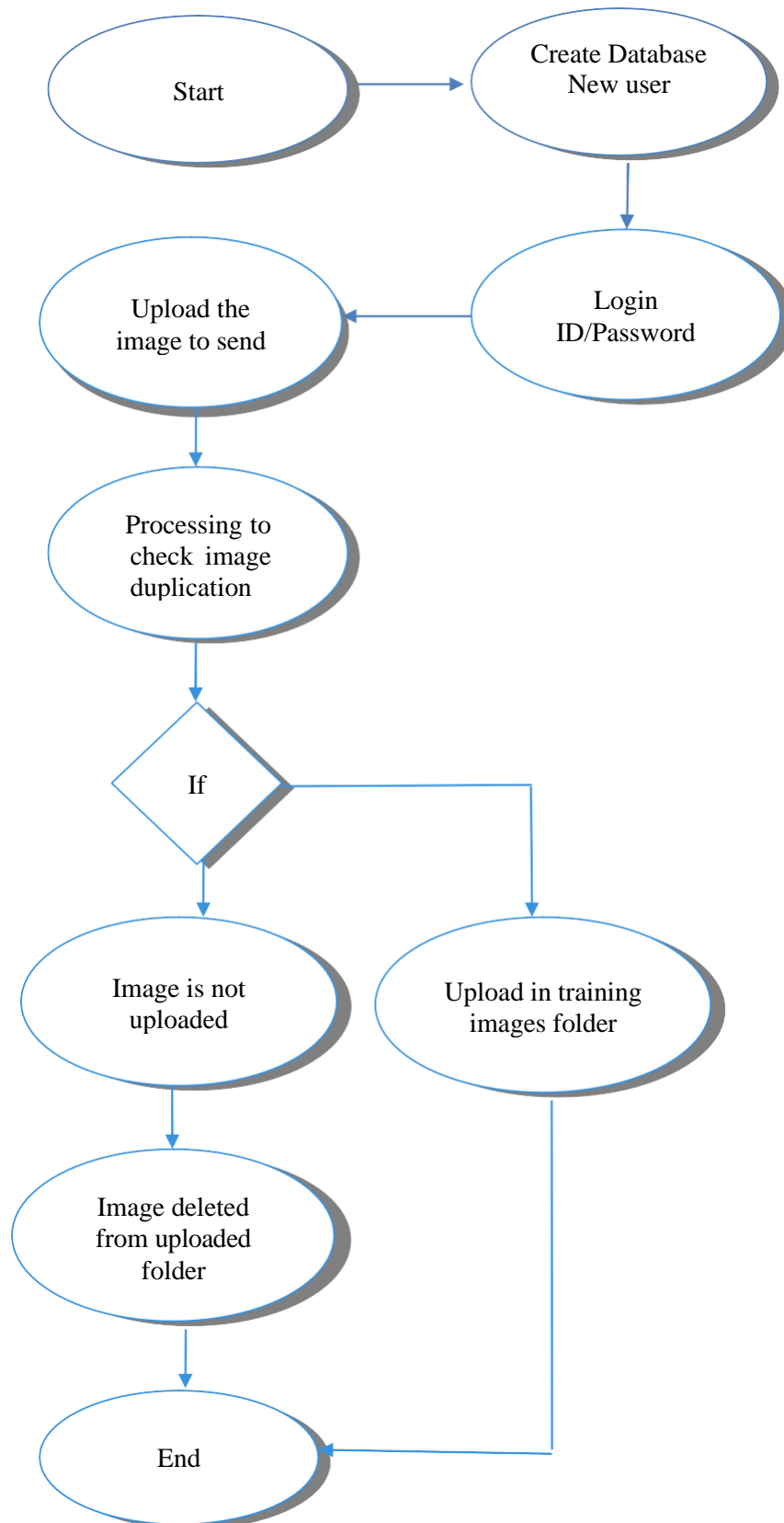


Figure 4.5 Application Flowchart

CHAPTER 5

IMPLEMENTATION AND RESULT

5.1 IMPLEMENTATION:

5.1.1 Database Connectivity to JAVA Application:

To connect Java application with the MySQL database, we need to follow 5 following steps.

- Driver class: The driver class for the mysql database is `com.mysql.jdbc.Driver`.
- Connection URL: The connection URL for the mysql database is `jdbc:mysql://localhost:8080/DataDeduplication` where `jdbc` is the API, `mysql` is the database, `localhost` is the server name on which mysql is running, we may also use IP address, 8080 is the port number and `DataDeduplication` is the database name.
- Username: The default username for the mysql database is `root`.
- Password: It is the password given by the user at the time of installing the mysql database. In project, we are going to use `root` as the password.

To connect to MySQL from Java, we have used the JDBC driver from MySQL. The MySQL JDBC driver is called MySQL Connector/J. JDBC provides an abstraction layer between Java applications and database servers, so that an application's code does not need to be altered in order for it to communicate with multiple database formats. Rather than connecting to the database directly, the applications send requests to the JDBC API, which in turn communicates with the specified database through a driver that converts the API calls into the proper dialect for the database to understand.

5.1.2 Web Scraping:

Web scraping refers to the process of extracting a significant amount of information from a website using scripts or programs. Such scripts or programs allow one to extract data from a website, store it and present it as designed by the creator. We will extract the information which was entered by the user from webpage to create user's own database for handling the images and texts. The time taken to extract information from a particular source is significantly reduced as compared to manually copying and pasting the data if we use web scraping.

5.1.3 Image Comparison:

After successful creation of user database, the user will upload images to store on database. Each time the user uploads the new image, the application will calculate average RGB (Red Green Blue) value of an image. The RGB value can be find out by using method “getRGB()”. Simultaneously, by using K- Means Clustering algorithm the image gets divided into multiple chunks.

Both shape of an image and average RGB value will get compared with the images present in database. Firstly, the application will compare the values and list out the images which have the value near to the value of a current uploaded image. Secondly, clustering algorithm will follow the output of a segmentation stage, which usually is raw pixel data, constituting either the boundary of a region (i.e., the set of pixels separating one image region from another) or all the points in the region itself. This region of a current uploaded will be compared with other image’s region. Finally, if the application does not found any duplicate image then the image will be stored in user’s database else the image will be deleted.

5.1.4 K-Means Clustering Algorithm for Segmentation:

K-Means clustering algorithm is an unsupervised algorithm and it is used to segment the interest area from the background. It clusters, or partitions the given data into K-clusters or parts based on the K-centroids. The algorithm is used when you have unlabeled data (i.e. data without defined categories or groups). The goal is to find certain groups based on some kind of similarity in the data with the number of groups represented by K. The objective of K-Means clustering is to minimize the sum of squared distances between all points and the cluster center.

The diagram shows the objective function formula for K-Means clustering: $J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$. Annotations include: 'number of clusters' pointing to k , 'number of cases' pointing to n , 'case i ' pointing to $x_i^{(j)}$, 'centroid for cluster j ' pointing to c_j , and 'Distance function' pointing to the term $\|x_i^{(j)} - c_j\|^2$. An arrow labeled 'objective function' points to the entire equation.

- **Steps in K-Means algorithm:**

1. Choose the number of clusters K.
2. Select at random K points, the centroids(not necessarily from your dataset).
3. Assign each data point to the closest centroid → that forms K clusters.
4. Compute and place the new centroid of each cluster.
5. Reassign each data point to the new closest centroid. If any reassignment took place, go to step 4, otherwise, the model is ready.

- **Key Features of k-means Clustering Algorithm:**

1. It is very smooth in terms of interpretation and resolution.
2. For a large number of variables present in the dataset, K-means operates quicker than Hierarchical clustering.
3. While redetermining the cluster center, an instance can modify the cluster.
4. K-means reforms compact clusters.
5. It can work on unlabeled numerical data.

- **Limitations with K-means:**

1. Sometimes, it is quite tough to forecast the number of clusters, or the value of k.
2. The output is highly influenced by original input, for example, the number of clusters.
3. An array of data substantially hits the concluding outcomes.
4. In some cases, clusters show complex spatial views, then executing clustering is not a good choice.
5. Also, rescaling is sometimes conscious, it can't be done by normalization or standardization of data points, the output gets changed entirely

5.2 APPLICATION OF IMAGE PROCESSING:

- **Object Extraction from an Image or Video**

Object Extraction is a closely related issue with the segmentation process. Image Segmentation is a process of dividing an image into sub partition based on some characteristics like color, intensity etc. The main goal of object extraction is to change the representation of an image into something more meaningful. To extract an object from the image first we have to segment the entire image. User select the region as background and foreground by using the markers and then the algorithm will segment the image and the foreground region will be extracted from the image. In future we can also be able to extract the required object from video with further improvement of this technology.

- **Remote Sensing**

For this application, sensors capture the pictures of the earth's surface in remote sensing satellites or multi – spectral scanner which is mounted on an aircraft. These pictures are processed by transmitting it to the Earth station. Techniques used to interpret the objects and regions are used in flood control, city planning, resource mobilization, agricultural production monitoring, etc.

- **Medical Imaging**

Medical image processing tools are playing an increasingly important role in assisting the clinicians in diagnosis, therapy planning and image-guided interventions. Accurate, robust and fast tracking of deformable anatomical objects such as the heart is a crucial task in medical image analysis.

- **Face Detection and Face Recognition**

Face detection and Face Recognition is widely used in computer vision task. We noticed how Facebook detects our face when you upload a photo .This is a simple application of object detection that we see in our daily life.

Face detection can be regarded as a specific case of object-class detection. In object-class detection, the task is to find the locations and sizes of all objects in an image that belong to a given class. Examples include upper torsos, pedestrians, and cars.

Face detection is a computer technology being used in a variety of applications that identifies human faces in digital images. Face recognition describes a biometric technology that goes way beyond recognizing when a human face is present. It actually attempts to establish whose face it is.

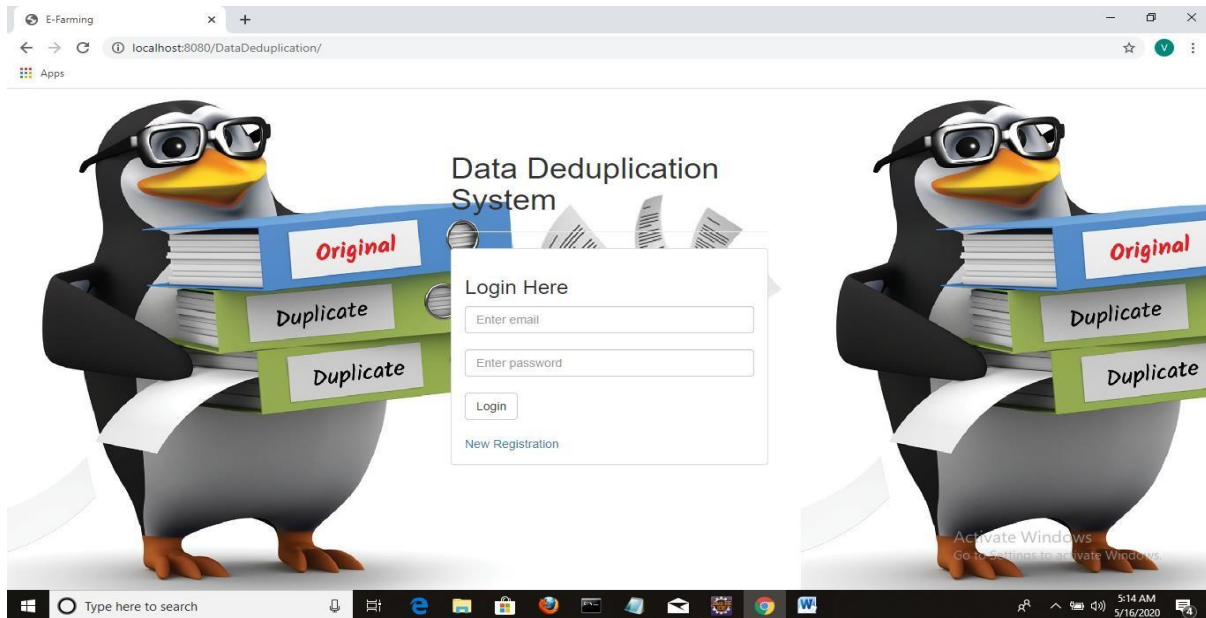
- **Defense surveillance**

Aerial surveillance methods are used to continuously keep an eye on the land and oceans. This application is also used to locate the types and formation of naval vessels of the ocean surface. The important duty is to divide the various objects present in the water body part of the image. The different parameters such as length, breadth, area, perimeter, compactness are set up to classify each of divided objects. It is important to recognize the distribution of these objects in different directions that are east, west, north, south, northeast, northwest, southeast and south west to explain all possible formations of the vessels. We can interpret the entire oceanic scenario from the spatial distribution of these objects.

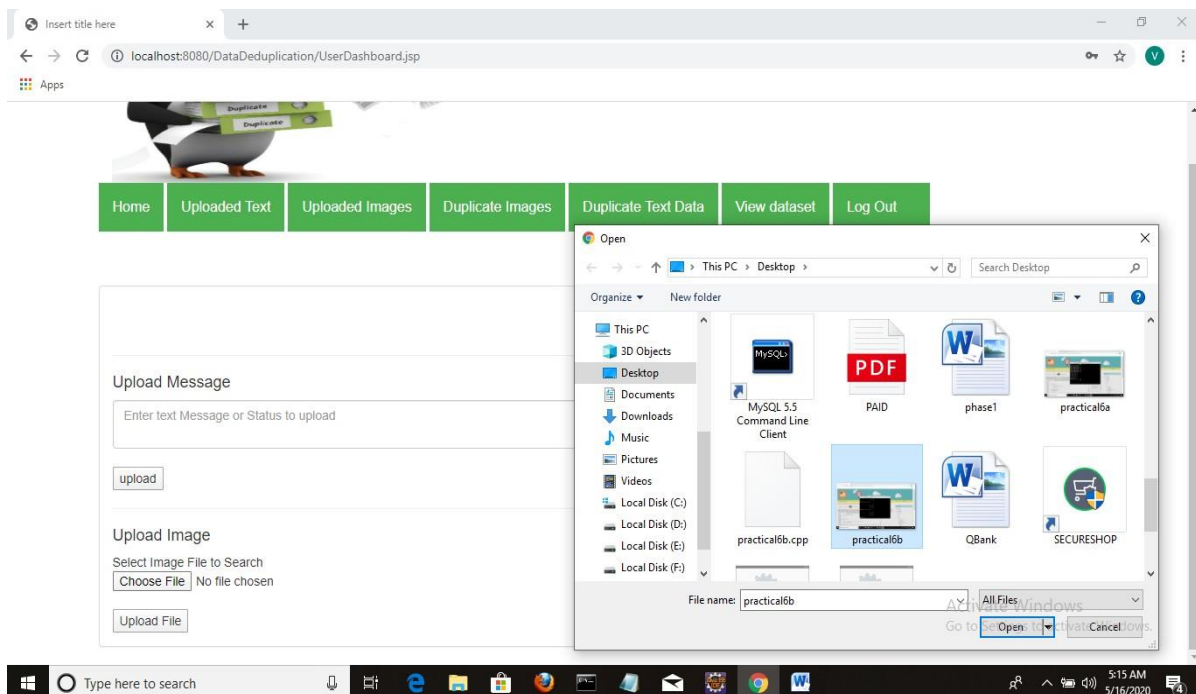
5.3 RESULTS:

5.3.1 : Checking for Duplicate Image:

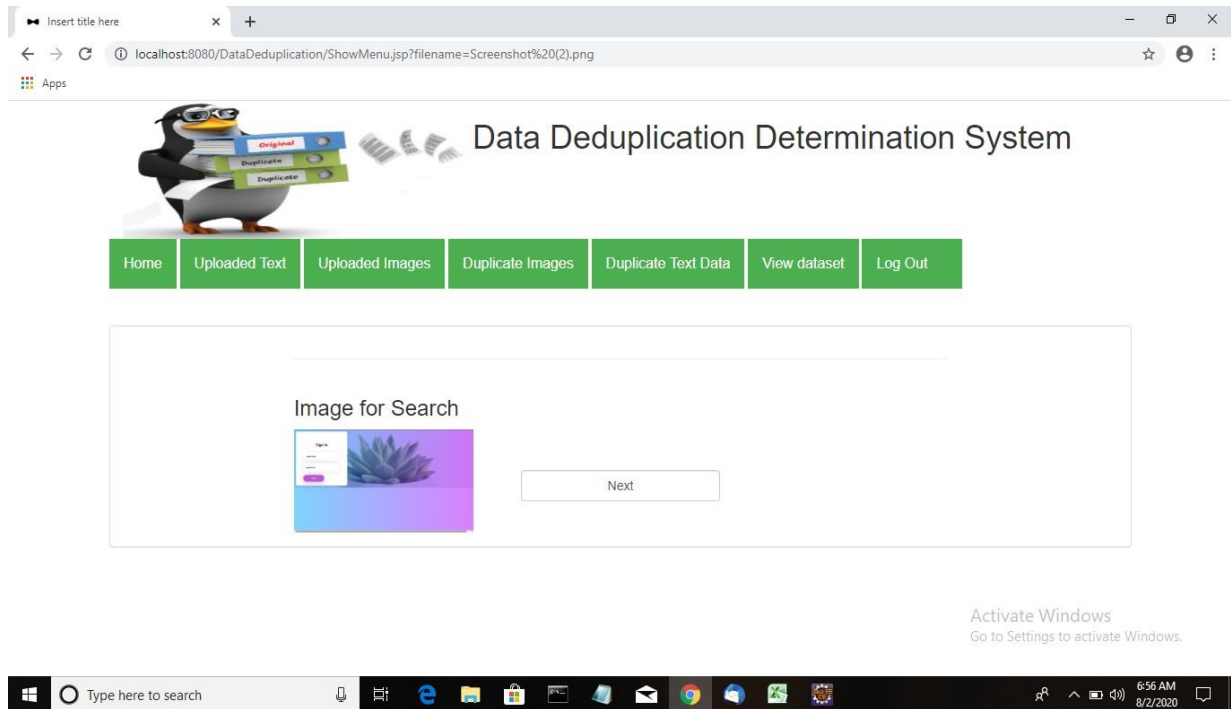
Screenshot 1: Login and Registration



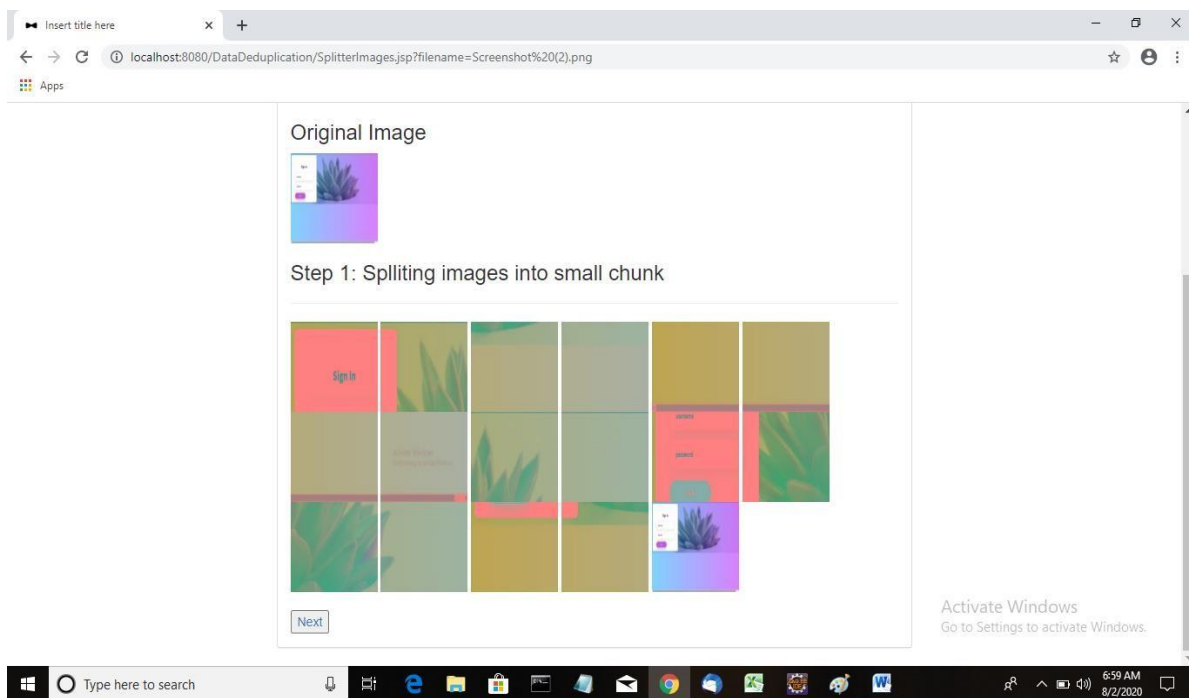
Screenshot 2: Options for Uploading Image or Text

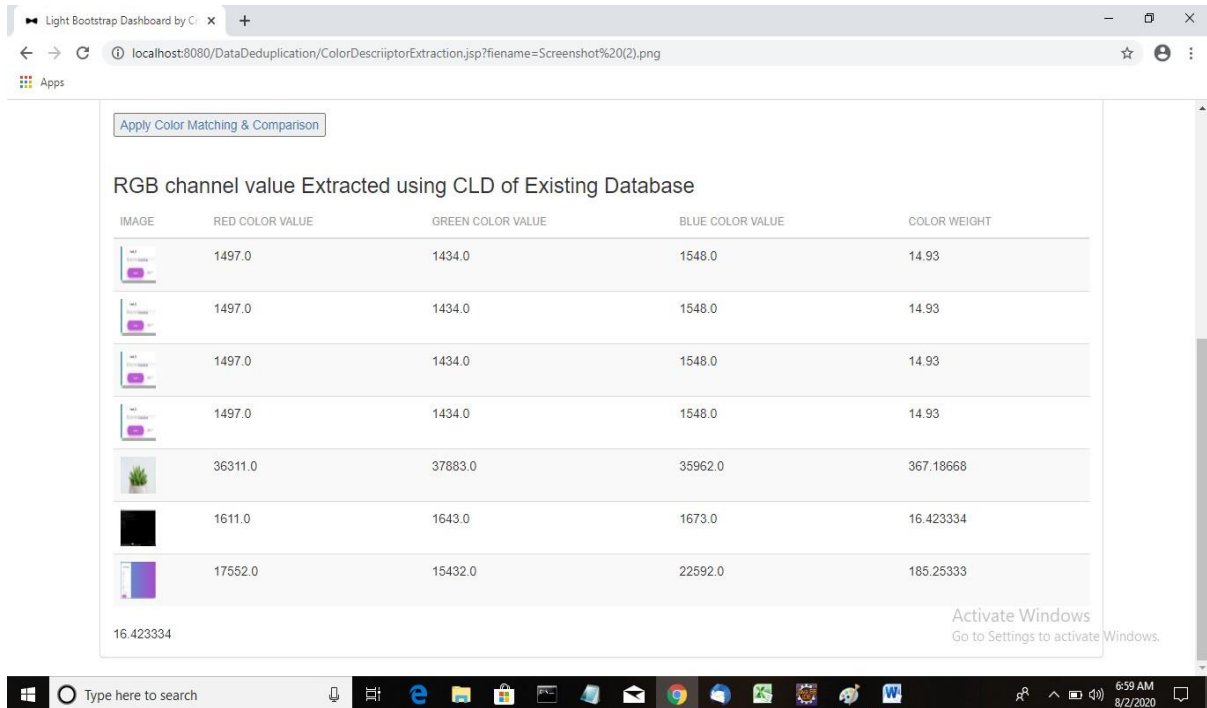


Screenshot 3: Image Upload










Screenshot 4: Segmentation Process



Screenshot 5: Previously Calculated RGB Values of Images


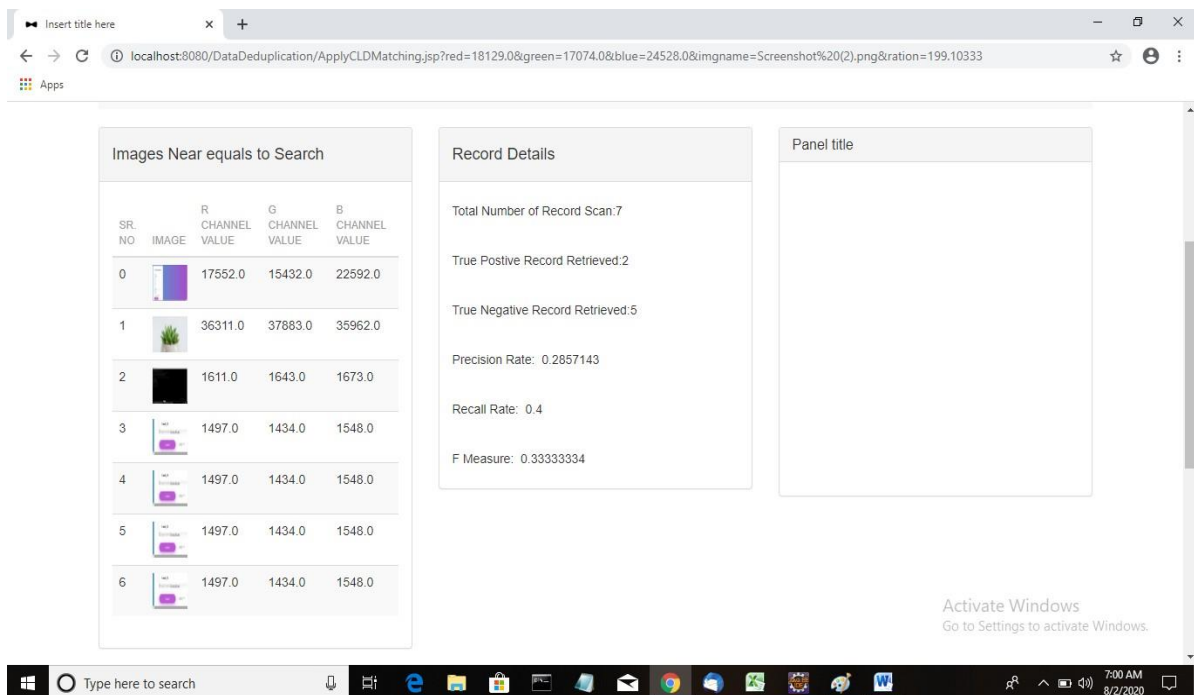
Apply Color Matching & Comparison

RGB channel value Extracted using CLD of Existing Database

IMAGE	RED COLOR VALUE	GREEN COLOR VALUE	BLUE COLOR VALUE	COLOR WEIGHT
	1497.0	1434.0	1548.0	14.93
	1497.0	1434.0	1548.0	14.93
	1497.0	1434.0	1548.0	14.93
	1497.0	1434.0	1548.0	14.93
	36311.0	37883.0	35962.0	367.18668
	1611.0	1643.0	1673.0	16.423334
	17552.0	15432.0	22592.0	185.25333

16.423334








Activate Windows
Go to Settings to activate Windows.

Screenshot 6: Comparing Images in Database


Insert title here

localhost:8080/DataDeduplication/ApplyCLDMatching.jsp?red=18129.0&green=17074.0&blue=24528.0&imgname=Screenshot%20(2).png&rlation=199.10333

Images Near equals to Search

SRL NO	IMAGE	R CHANNEL VALUE	G CHANNEL VALUE	B CHANNEL VALUE
0		17552.0	15432.0	22592.0
1		36311.0	37883.0	35962.0
2		1611.0	1643.0	1673.0
3		1497.0	1434.0	1548.0
4		1497.0	1434.0	1548.0
5		1497.0	1434.0	1548.0
6		1497.0	1434.0	1548.0

Record Details

Total Number of Record Scan: 7

True Positive Record Retrieved: 2

True Negative Record Retrieved: 5

Precision Rate: 0.2857143

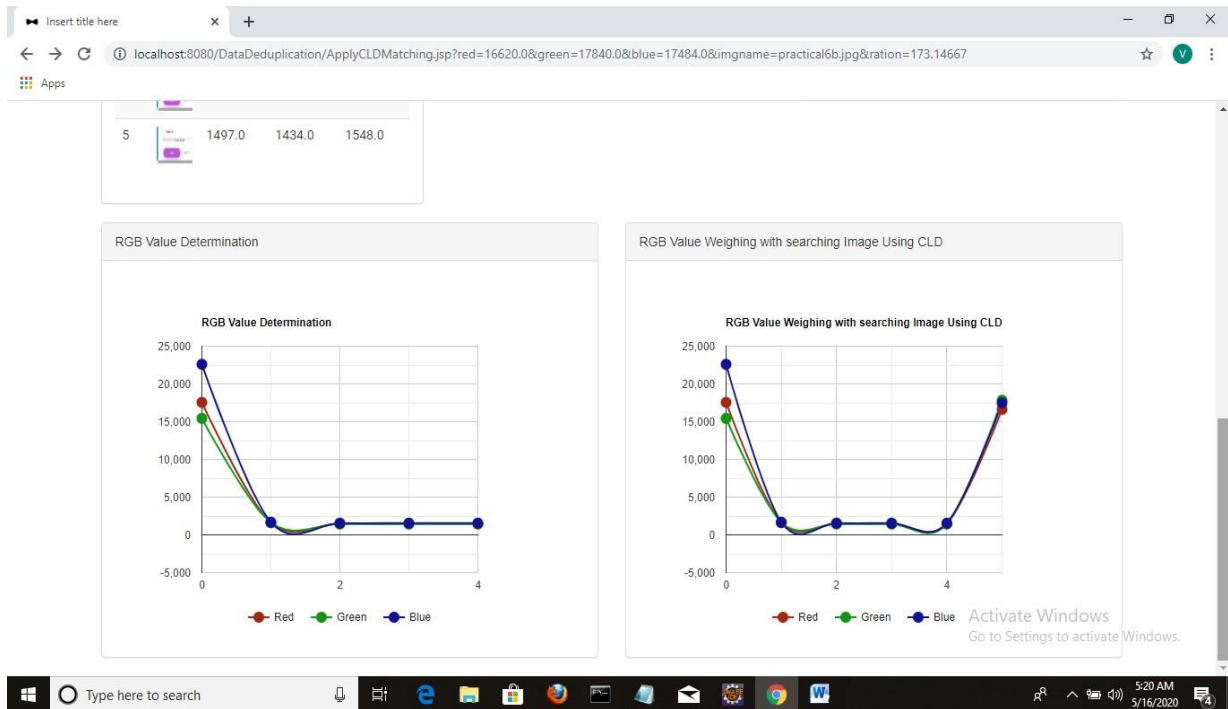
Recall Rate: 0.4

F Measure: 0.33333334

Panel title





Activate Windows
Go to Settings to activate Windows.

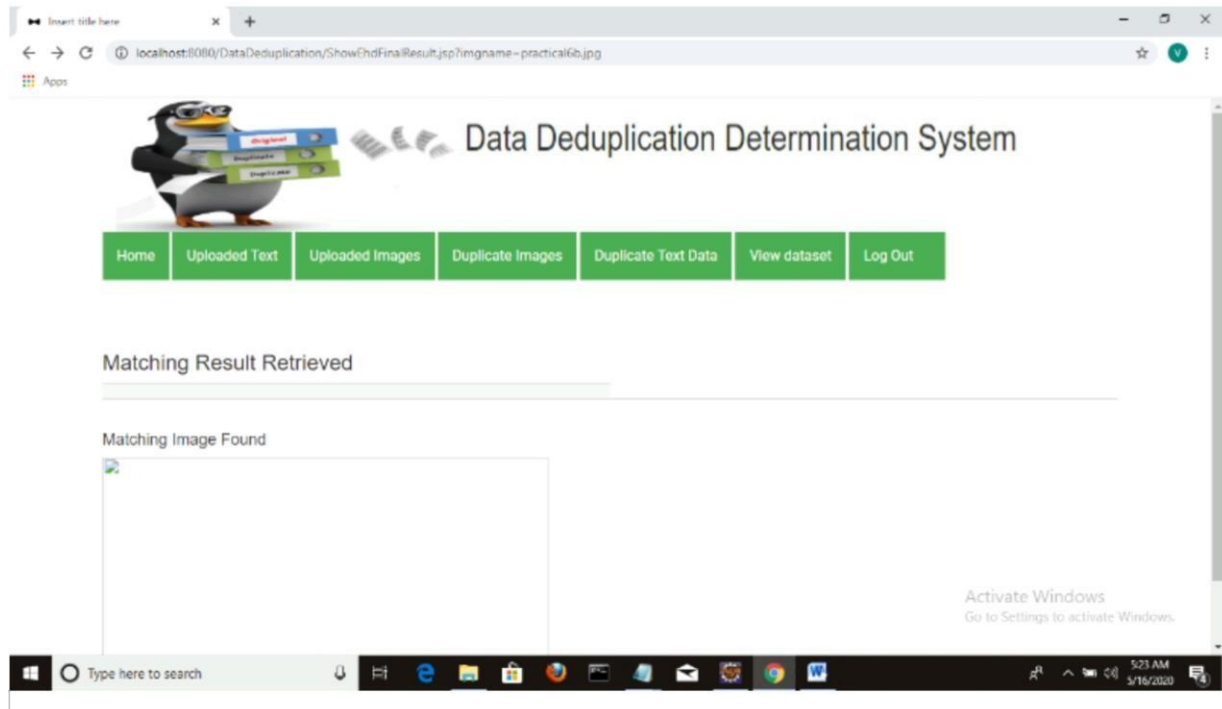
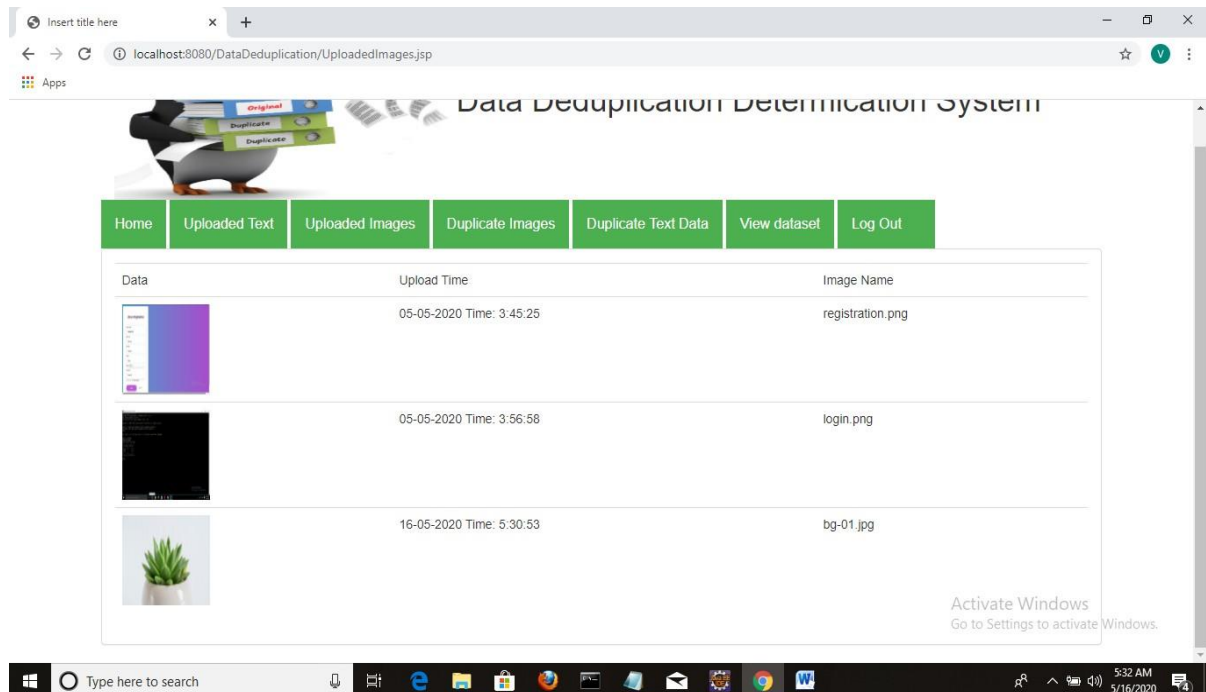
Screenshot 7: Comparison Graph of RGB values



Screenshot 8: Duplicate Copy Detected

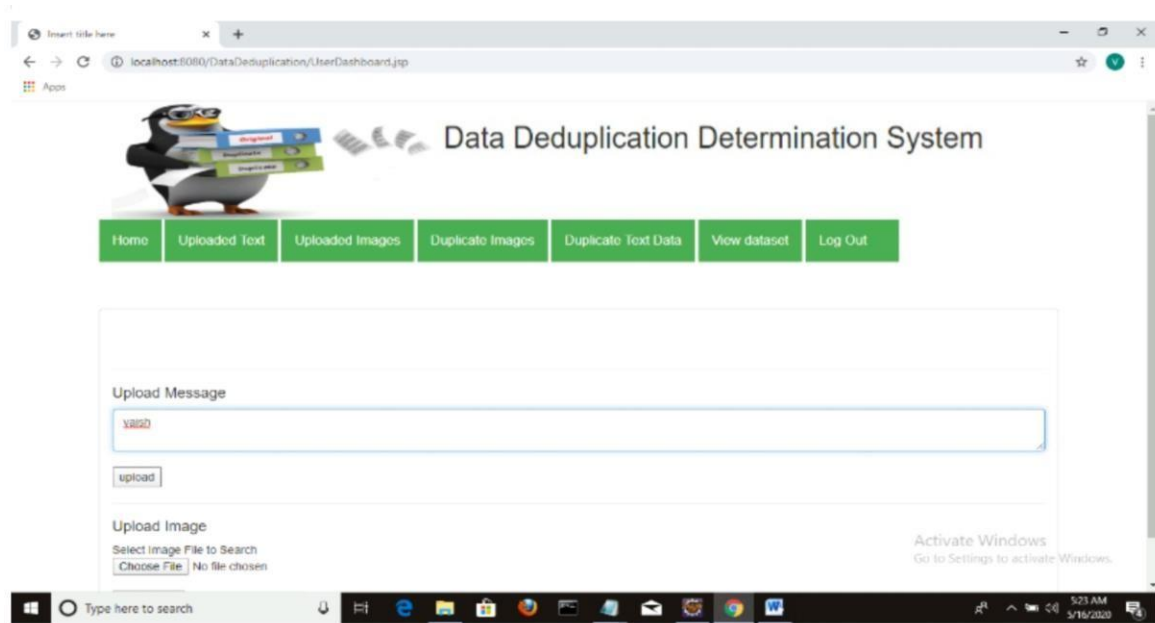
Check For Duplicate

IMAGE	IMAGE NAME	THRESHOLD
	12 - Copy (2).jpg	0.0
	12 - Copy (3).jpg	0.0
	12 - Copy.jpg	0.0
	12.jpg	0.0

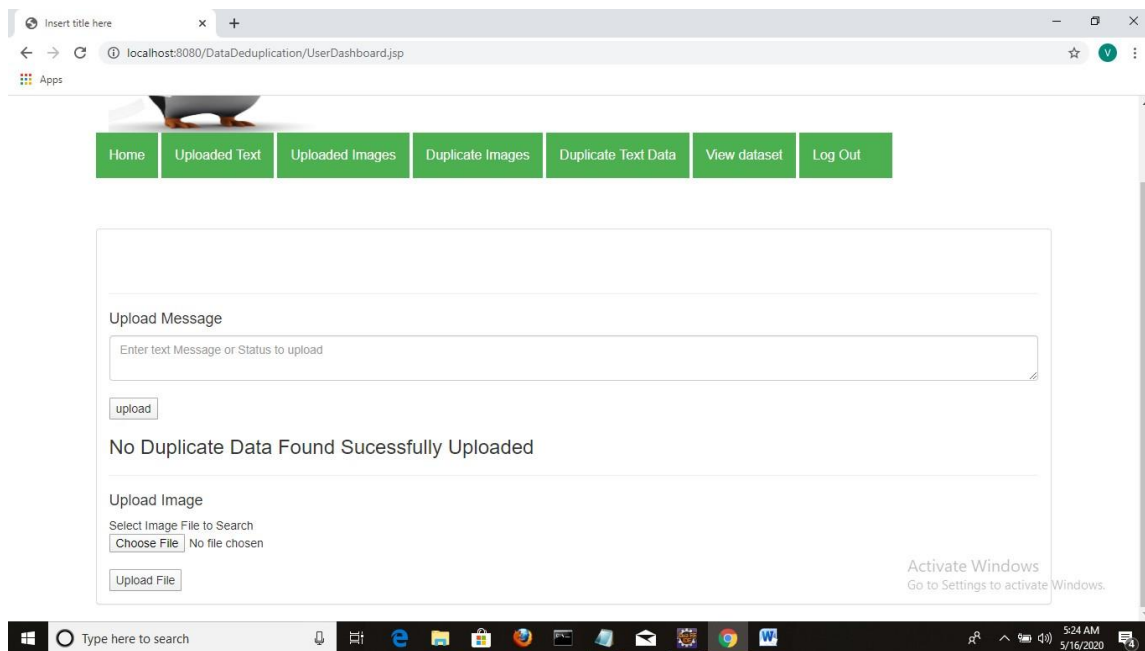
Screenshot 9: Showing the Result**Screenshot 10: Previously Uploaded Images with Time**

5.3.2: Checking for Duplicate Texts:

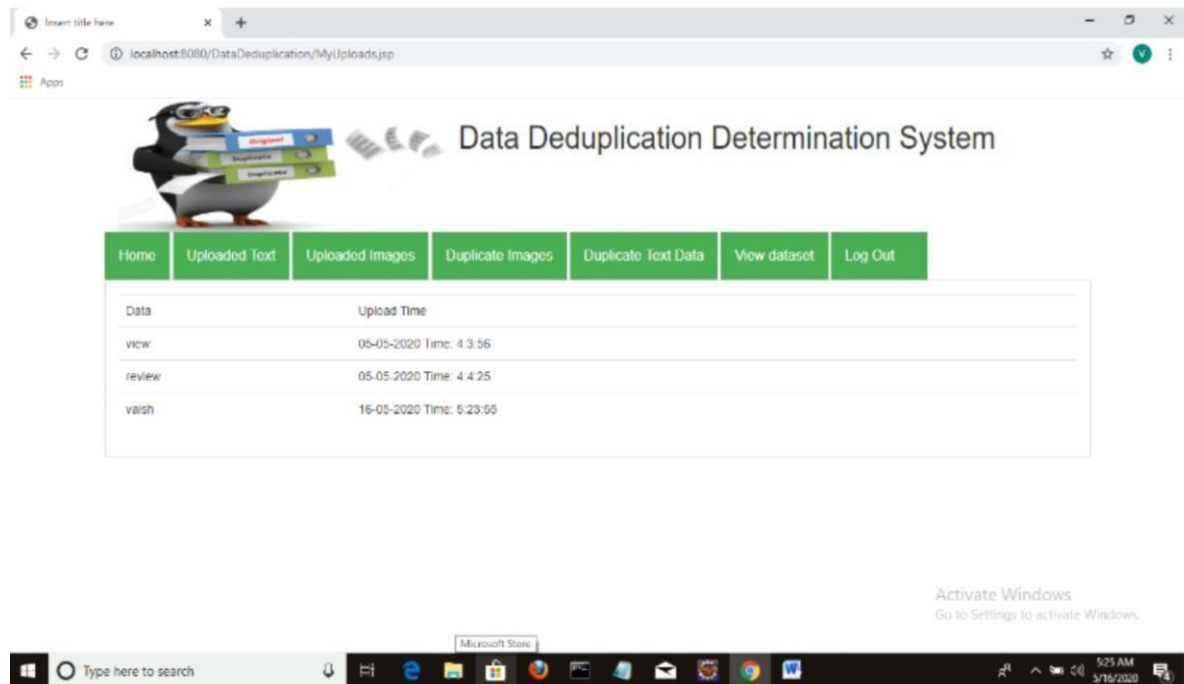
Screenshot 11: Uploading Text



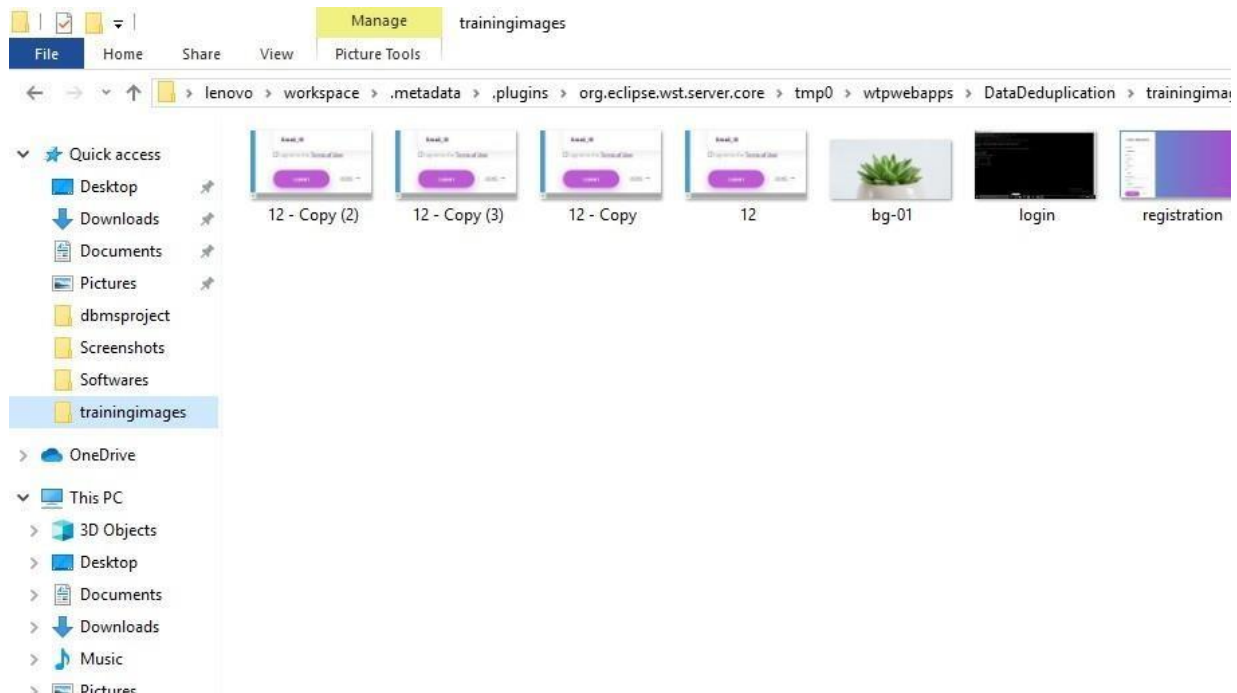
Screenshot 12: Result (No Duplicate data)



Screenshot 13: Previously Uploaded Texts with Time



Screenshot 14: Training Data Set



CHAPTER 6

CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

Data deduplication, an efficient approach to data reduction, has gained increasing attention and popularity in large-scale storage systems due to the explosive growth of digital data. It eliminates redundant data at the file or sub file level and identifies duplicate content by its cryptographically secure hash signature (i.e., collision-resistant fingerprint), which is shown to be much more computationally efficient than the traditional compression approaches in large-scale storage systems. The space savings that you can gain from Data Deduplication depend on the dataset or workload on the volume. Datasets that have high duplication could see optimization rates of up to 95%. This report presents scalable methods to detect duplicate data in user database and delete them to use the memory more effectively.

While deleting the images by using data duplication, we have used digital image processing technique. Digital image processing is a fascinating field, and is rightly seeing a ton of traction in commercial, as well as research applications. Understanding the images given by user and processing them to compare with others by using multiple processing algorithms available is a crucial task. This project will compare RGB values and finds out the identical data by using proper clustering algorithm for segmentation.

Choice of a right image processing method and algorithm is crucial and depends on the problem scenario one is trying to solve and the set-up. Digital image processing is backbone of many practical applications in fields like computer vision, object detection, etc. By use of digital image processing we can extract more information easily and use that information for future use.

6.2 FUTURE SCOPE

6.2.1 Future in Data Deduplication:

It may seem like data deduplication technology for backup has been around forever, but there are plenty of companies that have yet to add deduplication to their backup operations. The technology can lead to a significant reduction in required storage space, especially in situations where redundancy is high. As a result, data deduplication has firmly established itself in the backup market. But not every data center uses deduplication. For example, Storage magazine's most recent Purchasing Intentions survey found that more than 60% of data centers haven't added data deduplication technology to their backup operations. The level of resistance to deduplication may come as a surprise to many in the storage industry. While it appears to be a maturing technology and the term "deduplication" is so commonly used, it's easy to assume the technology is in use everywhere.

6.2.2 Future of Image Processing

We all are in midst of revolution ignited by fast development in computer technology and imaging. Against common belief, computers are not able to match humans in calculation related to image processing and analysis. But with increasing sophistication and power of the modern computing, computation will go beyond conventional, Von Neumann sequential architecture and would contemplate the optical execution too. Parallel and distributed computing paradigms are anticipated to improve responses for the image processing results.

The future of image processing will involve scanning the heavens for other intelligent life out in space. Also new intelligent, digital species created entirely by research scientists in various nations of the world will include advances in image processing applications. Due to advances in image processing and related technologies there will be millions and millions of robots in the world in a few decades time, transforming the way the world is managed. Advances in image processing and artificial intelligence⁶ will involve spoken commands, anticipating the information requirements of governments, translating languages, recognizing and tracking people and things, diagnosing medical conditions, performing surgery, reprogramming defects in human DNA, and automatic driving all forms of transport. With increasing power and sophistication of modern computing, the concept of computation can go beyond the present limits and in future, image processing technology will advance and the visual system of man can be replicated.

A wide research is being done in the Image Processing technique. Some of them are mentioned below:

- ❖ Cancer Imaging – Different tools such as PET, MRI, and computer aided detection helps to diagnose and be aware of the tumour.
- ❖ Brain Imaging – Focuses on the normal and abnormal development of brain, brain ageing and common disease states.
- ❖ Image processing – This research incorporates structural and functional MRI in neurology, analysis of bone shape and structure, development of functional imaging tools in oncology, and PET image processing software development.

REFERENCES

- [1] Xia, Wen & Jiang, Hong & Feng, Dan & Douglass, Fred & Shilane, Philip & Hua, Yu & Fu, Min & Zhang, Yucheng & Zhou, Yukun. (2016). A Comprehensive Study of the Past, Present, and Future of Data Deduplication. Proceedings of the IEEE. 1-30. 10.1109/JPROC.2016.2571298.
- [2] K. A. Abdul Nazeer and M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, In Proceedings of the World Congress on Engineering, London, WCE, vol. 1, July (2001).
- [3] Madhu Yedla, Srinivasa Rao Pathakota and T. M. Srinivasa , Enhanced K -means Clustering Algorithm with Improved Initial Center, In International Journal of Science and Information Technologies, vol. 1(2), pp. 121-125, (2010).
- [4] N. Akhtar, N. Agarwal and A. Burjwal, "K-mean algorithm for Image Segmentation using Neutrosophy," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014, pp. 2417-2421, doi: 10.1109/ICACCI.2014.6968286.
- [5] Rafael C. Gonzalez and Richard E. Woods, "Digital Image Processing", 2nd ed., Beijing: publishing firm of industry, 2007.
- [6] T. Shraddha, K. Krishna, B.K.Singh and R. P. Singh, "Image Segmentation: A Review", International Journal of engineering science and Management analysis Vol. 1 Issue. 4 November 2012.
- [7] Qinlu He, Zhanhuai Li and Xiao Zhang, "Data deduplication techniques," 2010 International Conference on Future Information Technology and Management Engineering, Changzhou, 2010, pp. 430-433, doi: 10.1109/FITME.2010.5656539.
- [8] C. Tseng, J. Ciou and T. Liu, "A Cluster-Based Data De-duplication Technology," 2014 Second International Symposium on Computing and Networking, Shizuoka, 2014, pp. 226-230, doi: 10.1109/CANDAR.2014.22.
- [9] M. Jaehong, Y. Daeyoung and W. Youjip, "Efficient deduplication techniques for modern backup operation", IEEE Transactions on Computers, vol. 60, no. 8, pp. 824-840, Jun. 2011.

- [10] K K, Thyagarajan & Hasini, V & Sweetlin, Jenifer & Joshna, H. (2017). Near-Duplicate Image Identification using Pulse Coupled Neural Networks. International Journal of Advance Research in Computer Science and Management Studies. 5. 85-92.
- [11] Kalaiarasi, G. & K K, Thyagarajan. (2014). Clustering of near Duplicate Images in the Search Using Affine Transform and Hybrid Hierarchical K-means (HHK) Algorithm. WIT Transactions on Information and Communication Technologies. 60. 239.
- [12] Elmagarmid, Ahmed & Ipeirotis, Panos & Verykios, Vassilios. (2007). Duplicate Record Detection: A Survey. Knowledge and Data Engineering, IEEE Transactions on. 19. 1 - 16. 10.1109/TKDE.2007.250581
- [13]<https://towardsdatascience.com/introduction-to-image-segmentation-with-k-means-clustering-83fd0a9e2fc3?gi=51d72533cd51>
- [14]<https://community.spiceworks.com/topic/1999524-what-is-deduplication-definition-benefits-pros-and-cons-word-of-the-week>
- [15] Zheng, X., Lei, Q., Yao, R. et al. Image segmentation based on adaptive K-means algorithm. J Image Video Proc. 2018, 68 (2018). <https://doi.org/10.1186/s13640-018-0309-3>
- [16] Pallavi Purohit and Ritesh Joshi, A New Efficient Approach towards k-means Clustering Algorithm, In International Journal of Computer Applications, (0975-8887), vol. 65, no. 11, March (2013).
- [17] Alan Jose, S. Ravi and M. Sambath, Brain Tumor Segmentation using K-means Clustering and Fuzzy C-means Algorithm and its Area Calculation. In International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, issue 2, March (2014).
- [18] Madhu Yedla, Srinivasa Rao Pathakota and T. M. Srinivasa, Enhanced K-means Clustering Algorithm with Improved Initial Center, In International Journal of Science and Information Technologies, vol. 1(2), pp. 121–125, (2010).
- [19] Gonzalez, Rafael (2018). Digital image processing. New York, NY: Pearson.
- [20] Systems, C. US Secure Hash Algorithm 1 (SHA1). Retrieved September 2001