TEXT AND IMAGE PLAGIARISM DETECTION

A Project Work-1 Internal Review Report

Submitted in Partial Fulfillment for the Award of the Degree Of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

Submitted by

CHALAPA BALA MURALI KRISHNA	20B95A1203
SHAIK GOUSE MOHAMMED AHMED ALISHA	19B91A12G3
PANDURI SURYA TEJA	20B95A1214
VAJJIPARTHI SRINIVAS	20B95A1216

Under the esteemed guidance of

Sri K. Chandra Sekhar

M-Tech (IT)

Assistant Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

S.R.K.R ENGINEERING COLLEGE (AUTONOMOUS)

(Approved by AICTE, NewDelhi, Affiliated to JNTU University, Kakinada)

CHINNA AMIRAM :: BHIMAVARAM-534204

S.R.K.R ENGINEERINGCOLLEGE (A)

(Approved by AICTE, New Delhi, Affiliated to JNTUUNIVERSITY, KAKINADA}

CHINNA AMIRAM, BHIMAVARAM-534204

DEPARTMENT OF INFORMATION TECHNOLOGY

Certificate

This is to certify that the Project Work-1 Internal review report entitled "TEXT AND IMAGE PLAGIARISM DETECTION”, is bonafide work submitted by CHALAPA BALA MURALI KRISHNA (RegdNo:20B95A1203), SHAIK GOUSE MOHAMMED AHMED ALISHA (RegdNo:19B91A12G3), PANDURI SURYA TEJA (RegdNo:20B95A1214), VAJJIPARTHI SRINIVAS (RegdNo:20B91A1216) in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Information Technology during the Academic year 2022-2023.

HOD Signature

Dr. Bh. V. S. R. K Raju

Professor, HOD-IT Dept

GUIDE Signature

K. Chandra Sekhar

Assistant Professor

CERTIFICATION OF EXAMINATION

This to certify that I have examined the concept and hereby accord my approval of it as a Project Work-1 carried out and presented in a manner required for its acceptance on partial fulfillments for the award of the degree of BACHELOR OF TECHNOLOGY for which it has been submitted.

This approval does not necessarily endorse or accept every statement made opinion expressed or conclusions drawn as recorded in the Project Work-1 Internal Review report it only signifies the acceptance of the report for the purpose for which submitted.

Signature:

Project Guide	**Internal Examiner**	**HOD**

DECLARATION

This project work-1 Internal review report entitled "TEXT AND IMAGE PLAGARISM DETECTION" has been carried out by us in the partial fulfillment of the requirements for the award of the degree of B.TECH (IT), S.R.K.R Engineering College. We here by declare this project work/project report has not been submitted to any of the other university/Institute for the award of any other degree/diploma.

CHALAPA BALA MURALI KRISHNA	20B95A1203
SHAIK GOUSE MOHAMMED AHMED ALISHA	19B91A12G3
PANDURI SURYA TEJA	20B95A1214
VAJJIPARTHI SRINIVAS	20B95A1216

INDEX

TEXT AND IMAGE PLAGIARISM DETECTION

1.INTRODUCTION

1.1 INTRODUCTION ABOUT PROBLEM YOU ARE TRYING TO SOLVE:

Plagiarism in itself cannot be considered as a crime but as copyright violation. In the academics and other industries that are sensitive to copyright infringement, plagiarism is grave misconduct in integrity. The law cannot and usually will not punish plagiarism, but it is up to the institution on how to handle it once it happens Plagiarism detection is usually split into two which is text-based plagiarism detection and image-based plagiarism detection. For text-based plagiarism detection there are currently five techniques that is used most often in different fields. These techniques are Fingerprinting, String Matching, Bag of Words, Citation Analysis and Stylometry. String Matching is mostly used in computer science where it compares the documents words for words. Bag of words represents the documents in one or two vectors for comparison. Citation analysis is mainly used in scientific texts because it only compares the citation and reference of the documents. Stylometry check the author’s unique writing style for detection author’s ownership. For image-based plagiarism detection, there is no commonly used techniques like the text-based plagiarism detection, but they usually share the same processes and steps. When we say plagiarism checking or detection we usually mean checking only the text in the file or document for plagiarism. Most of the times when you check your documents or files for plagiarism through a plagiarism checker software they will check for images and then discard them Plagiarism basically means the wrongful stealing of an author’s work, thoughts, ideas, etc. and claiming it as your own original work. Plagiarism is considered as deceit and a breach of ethics. In academics, students that are caught with plagiarism are exposed to various level of penalties and punishment and may even lead to expulsion and we have to try to solve.

1.2 INTRODUCTION ABOUT EXISTING SYSTEM:

Plagiarism is the practice of copying someone else's work or ideas, and passing them off as one's own original work. Not only images but, architecture, flow diagram, UML diagrams, even snapshots of test results can be plagiarized. If the author has not mentioned the credit for the original author from where he/she copied the image then it is said to be plagiarized.

1.3 INTRODUCTION ABOUT PROPOSED SYSTEM:

The proposed work mainly focusses on finding the similarity between two images. Sample image is given as the reference and it is compared with the other image which is taken from any journal and comparison is done through histogram. Histogram is the best way to visualize the largest intensities of an image. It is used to find the problems which originate during image acquisition such as exposure, contrast etc. even a minute difference with the pixel is noticed by histogram

1.4 BRIEF INTRODUCTION ABOUT PLATFROM AND TECHNOLOGY USED AND WHY:

Python :

1. Python is currently the most widely used multi-purpose, high-level programming language.

2. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java.
3. Programmers have to type relatively less and indentation requirement of the language, makes them readable all the time.
4. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

MACHINE LEARNING: -

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

1.5 PURPOSE OF WORK:

The algorithm describes about the comparison between the images. In the initial stage the two images are loaded into the work space. Later the images are re-sized in-order to get the same size for both the images so that the result will be accurate when compared. The image is also compressed in order to improve the accuracy while comparing. Image is automatically saved in a folder after compression. To over- come the weakness of having images with different colors, the image is converted into grayscale image. In the next step the similar features of the images are detected and the comparison is done using image subtraction method where each pixel value of an image is compared with pixel value of another image and finally the result will be displayed.

1.6 SCOPE OF WORK:

In recent times, the use of internet has widely increased which is leading to easy opportunity of plagiarism, the proposed system will help in detecting the same. So the plagiarism detection will be very helpful in the future. Our system can also be used as ‘search by

2. LITERATURE SURVEY

2.1 SURVEY AND STUDY OF PUBLISHED LITERTURE ON THE ASSIGNED TOPIC :

[1] This paper gives a brief idea about classification, the classification is done based on language in the documents. Languages are classified as Mono-lingual and cross-lingual or multi-lingual. Mono-lingual plagiarism detection identifies and extracts texts from the document and detects language of same kind i.e English-English plagiarism. Cross-lingual or multi-lingual plagiarism detection also deals with identification and extraction of text from document and detects language of different kinds i.e English-Arabic plagiarism. [2]Shape-Based Plagiarism Detection for Flowchart Figures in Texts does pre-processing by determining the boundaries, edges, distance and the figures are stored in database by eliminating the text from the figures. The system takes the sample figure and pre-processes it to build the query vector that will be compared with the figure-document stored in the database, this will be the training phase. Then the test figures are given as input to the system and compares with the figures stored in the database. The result is the number of figures copied from the original paper. [3]In this paper, we compare Set A, B as two RGB images with same size, comparing A and B is to detect the same color of pixels with same location, the steps and algorithms. C is an image matrix from color matrix A subtracting color matrix B, then C=A-B, if the corresponding pixels of A and B have same color, then the RGB significance of corresponding pixels in image C should be 0, which means black, so the copied pixels between image A and B should be black. H is a set with black pixels extracted from image C, so all copied pixels between image A and B should be contained in H. As the images A, B may have the same background color, when comparing A and B, the part with same background color will be extracted to set H, therefore the therefore part must be eliminated recurring to the character of background color that it's usually monochrome. [4]The study of Histogram describes about the applications of how histograms can be used to know the properties of the image, enhancement of the image, to detect exposure saturation, brightness, gaps etc, and it also helps in thresholding. This paper also deals with Histogram stretching which determines the contrast of the images. Histogram sliding shows the intensity and brightness of the images. Histogram Equalization equalizes all the pixels of the image to one form which gives us the flat graph. [5]Flowchart Plagiarism Detection method uses area detection technique to detect plagiarism, the flowchart

images are given as input to the system which are pre- processed by detecting the edge using ’cannyedge detection’. For each shape in the image, the centroid and boundary is detected. Euclidean distance is calculated from centroid to boundary and a graph is generated. Then the generated graph is compared with the original image graph. The result is an alert displaying whether the image is plagiarized or not. This drawback of this approach is that it only works on flowchart images. [6]The paper ’Edge Detection Methods’ describes about the edge detection which an important feature extraction method, this method can be used to determine the lines in the images. The Author classifies different edge detection techniques like Sabel, Prewitt’s, Robert’s Cross, Laplacian of Guassian and Canny. Sample images are converted to grayscaleonwhichtheexperimentsareperformedonall the techniques. The comparison between these techniques is explained and concluded that Canny is best compared to all the techniques. [7]Content-Based Image Retrieval (CBIR) is a kind of feature extraction method which uses may contents of an image like shape, color, texture for representation and indexing of image.

3 PROBLEM STATEMENT

3.1 EXISTING SYSTEM

The existing methodology maybe sufficient for detecting plagiarism of images when the source and suspected image have not been rotated by a large margin, but in case of rotational changes the existing methodology will fail. The proposed methodology will ensure that even if the image is rotated plagiarism is detected if it has occurred or if an attack of rotational change has been made. Also the existing system is not efficient to detect plagiarism properly for different types of images. The proposed system will ensure that by using adaptive threshold values. The algorithm makes sure that the matching time of the images is less by reducing the search field by a significant factor each time the refinement is done

3.2 PROBLEM DEFINITION:

A research problem is defined as an area of concern that requires a meaningful understanding of a specific topic, a condition, a contradiction, or a difficulty. So what is research problem? A research problem means finding answers to questions or strengthening existing findings to bridge the knowledge gap to solve problems

3.3 PROBLEM STATEMENT:

The Indeed Editorial Team comprises a diverse and talented team of writers, researchers and subject matter experts equipped with Indeed's data and insights to deliver useful tips to help guide your career journey. A problem statement addresses issues in a timely and efficient manner. They help professionals break down complex situations into tangible goals that they can then communicate throughout an organisation. In every workplace, problems are inevitable. Thus, a problem statement is an effective tool to put into practice so that employees recognise issues before they disrupt multiple functions of the business. In this article, we discuss what a problem statement is, why they are important, how to write one and provide a comprehensive template and example for your reference**.**

4. SYSTEM ARCHITECTURE:

4.1 BLOCK DIAGRAM :

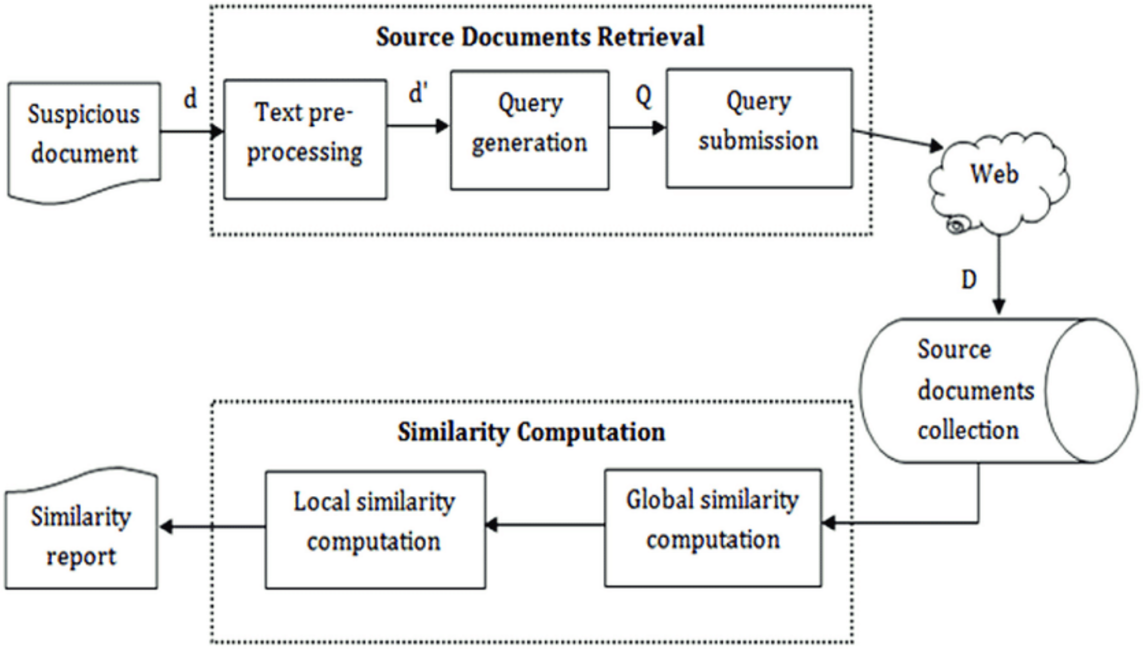

Fig. 4.1(a): Architecture diagram of text and image plagarism detection

4.2 MODULES EXPLANATION :

1.NEW USER SIGNUP

Firstly, user will register in to Application. It helpful to login into Application with username and password.

2.LOGIN

User will login into Application through username and password**.**

3.UPLOAD SOURCE FILE:

Folder is created into Upload Source Files' link to load all files from corpus folder

4.UPLOAD SUSPICIOUS FILE:

To load suspicious file and get result. user will upload file to Upload Suspicious files the result is execute. LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result.

5.UPLOAD SOURCE IMAGE:

In this module from all database images histogram will be calculated and store in array and whenever we upload new test image then both histogram will get matched.

6.UPLOAD SUSPICOUS IMAGE:

we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected. histogram pixel matching score is 15173 out of 40000 pixels so image is not plagiarized and now upload image from "images" folder and see result. we can both original and uploaded image histogram is matching 100% so plagiarism is detected and now get below result. histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result

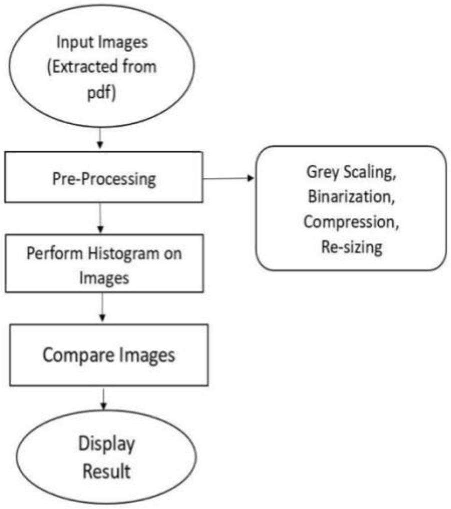
Input Images (Extracted from pdf)
Pre-Processing
Grey Scaling, Binarization, Compression, Re-sizing
Perform Histogram on Images
Compare Images
Display Result

Fig. 4.1(b) : Flow diagram of text and image plagiarism detection

5. PRELIMINARY ANALYSIS

5.1 BREIF ABOUT INPUT DATA:

Plagiarism in research is being debated more than ever before. There have been considerable harms to research as a consequence of web conditions and the ability to do complicated and intelligent searches in a short period of time

5.2 TYPE OF ANALYSIS DOING ON DATA:

The first step in any plagiarism analysis is preparing the materials to be checked for the software. Plagiarism checkers require documents with machine-readable text (such as Docx, RTF, etc.) and work best if that text is cleaned up and formatted correctly.

After preparing the documents, the next step is to actually perform the automated analysis.

Histogram

Sample image is given as the reference and it is compared with the other image which is taken from any journal and comparison is done through histogram. Histogram is the best way to visualize the largest intensities of an image. It is used to find the problems which originate during image acquisition such as exposure, contrast etc. even a minute difference with the pixel is noticed by histogram.

FIVE MODULUS METHOD

In most of images, there is a common feature which is the neighboring pixels are correlated. Therefore, finding a less correlated representation of image is one of the most important tasks. One of the basic concepts in compression is the reduction of redundancy and Irrelevancy. This can be done by removing duplication from the image. Sometime, Human Visual System (HVS) can not notice some parts of the signal, i.e. omitting these parts will not be noticed by the receiver. This is called as Irrelevancy.

Also, for bi-level images, the principle of image compression tells us that the neighbours of a pixel tend to be similar to the pixel. According to [2], this principle can be extended as that if the current pixel has any colour (black or white), then pixels seen in the past or future of the same color tend to have the same neighbours.

Hence, our proposed technique which is called Five Modulus Method (shortly FFM) is consists of dividing the image into blocks of 8×8 pixels each. Clearly, we know that each pixel is a number between 0 to 255 for each of the Red, Green, and Blue arrays. Therefore, if we can transform each number in that range into a number divisible by 5, then this will not affect the Human Visual System (HVS). Mathematically speaking, any number divided by 5 will give a remainder ranges from 0-4 (e.g., 15 mod 5 is 0, 17 mod 5 is 2, 201 mod 5 is 1, 187 mod 5 is 2 and so on). Here, we have proposed a new formula to transform any number in the range 0-255 into a number that when divided by 5 the result is always lying between 0-4.

Therefore, the pixels 200, 201, and 202 are the same for the human eye. Hence, a novel algorithm have been proposed to transform each pixel in the range 0-255 into the following numbers 0,5,10,15,20,25,30,35,40,...,200, 205,210,215,...,250, 255, (i.e. multiples of 5). Actually, any number in the range 0-4 (which is the remainder of dividing 0-255 by 5) can be transformed as follows 0→(same pixel), 1→(-1), 2→(-2), 3→(+2), 4→(+1). The algorithm can be described as:

Algorithm

Step1: Input images (Input two images into the work space).

Step 2: Images should be re-sized

(As we are comparing the image, the image should not vary in length and width, hence it should be re-sized).

Step 3: Compare the images and find the similar features among them.

Step 4: Display the result either the images are plagiarized or not.

The algorithm describes about the comparison between the images. In the initial stage the two images are loaded into the work space. Later the images are re-sized in-order to get the same size for both the images so that the result will be

accurate when compared. The image is also compressed in order to improve the accuracy while comparing. Image is automatically saved in a folder after compression. To over- come the weakness of having images with different colours, the image is converted into grayscale image. In the next step the similar features of the images are detected and the comparison is done using image subtraction method where each pixel value of an image is compared with pixel value of another image and finally the result will be displayed.

5.3 EXPECTED OUTCOME:

Once run through the software, what we have isn't a report of all of the plagiarism in the work, but of all the duplicative text that the checker found.

In the results phase we saved set of images which are extracted from the pdf. Sample image will be given as reference image and the other image is the one which needs to be compared. When the images are same, the histogram shows the similarity between the images. And if the images are not same it displays the variations in pixel through histogram.

6. FEASIBILITY STUDY

WORKING OUT A PRELIMINARY APPROACH TO THE PROBLEM RELATION TO THE ASSIGNED TOPICS

6.1 TECHNICAL FEASIBILTY :

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

6.2 OPERATION FEASIBILITY :

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

6.3 ECONOMICAL FEASIBILITY :

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased

7. SUMMARY OF PROJECTS

Plagiarism is itself cannot be considered as a crime but as copyright violation. This project is all about text and image plagiarism checker and in this project we are trying to solve the plagiarism problem by using Machine learning. The Suspicious document is given as a input to the system and the system performs analysis on the input by making Histogram count and try to find the similarity with original document that already exists in the Database and if Similarity occurs it shows that the document is contains plagiarism.

8.REFERENCES

1. Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, ”Understand- ing Plagiarism Linguistic Patterns, Textual Features, and Detection Methods”, IEEE, Vol:42, Issue:2, PP:133-149,2012.
2. Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim,” Shape-Based Plagiarism Detection for Flowchart Figures in Texts”, International Journal of Computer Science Information Tech- nology , Vol:6, No:1,2014.
3. Wang Wen, Wang Yanb and Li Bingbing , ”Research on Plagiarism IdentificationofDigitalImages”,IEEE,2010.
4. Harpreet Kaur and Neelofar Sohi, ”A Study for Applications of Histogram in Image Enhancement”, The International Journal of Engineering and Science (IJES), Vol:6, Issue:6, PP:59-63,2017.
5. Jithin S Kuruvila, Midhun Lal V L, Rejin Roy, Tomin Baby, Sangeetha Jamal and Sherly K K, ”Flowchart Plagiarism Detection System: An Image Processing Approach”, 7th International Conference on Advances in Computing Communications,2017.
6. Joshi, M., & Khanna, K. A Similarity Measure Analysis Based Improved Approach For Plagiarism Detection.
7. Ghassan Mahmoudhusien Amerand Dr. Ahmed Mohamed Abushaala, ”Edge Detection Methods”, IEEE,2015.

8. Reshma Chaudhari and A.M Patil, ”Content Based Image Retrieval Using Color and Shape Features”, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol:1, Issue:5,2012.
9. Prajakta Ovhal, ” Detecting Plagiarism in Images”, 2015 International Conference on Information Processing(ICIP),2015.
10. Firas A. Jassim and Hind E. Qassim,” Five Modulus Method for Image Compression”, An International Journal (SIPIJ), Vol:3, No:5,2012.
11. Minh Anh Nguyen, ”Results Review of Detecting of Human Errors Algorithm for image files”.
12. Akshay S, ”Single Moving Object Detection and Tracking Using Horn Schunck Optical Flow Method”, International Journal of Applied Engineering Research, Vol:10, No:11,2015.
13. Akshay S, Apoorva P”Segmentation and classification of FMM compressed retinal images using watershed and canny segmentation and support vector machine”,2017 International Conference on Communication and Signal Processing(ICCSP),2018.

REFERENCES LINKS:

https://www.researchgate.net/publication/334226542_Image_Plagiarism_Detection_using_Compressed_Images

http://gradivareview.com/gallery/grj%203676.pdf

https://ieeexplore.ieee.org/document/7959317

https://github.com/topics/plagiarism-detection?o=asc&s=stars

REFERENCES BOOKS:

1. Stop Plagiarism: A Guide To Understanding and Prevention by Vibiana Bowman Cvetkovic Ed.

2**.** Preventing Plagiarism: Tips and Techniques by Laura Hennessey DeSena Provides strategies for identifying, combating, and preventing plagiarism.

3. The Little Book of Plagiarism by Richard A. Posner (2007).