Image Plagiarism Detection using Compressed Images

Akshay S, Chaitanya B N, Rishabh Kumar

Abstract*: *Image plagiarism is stealing of another's work and passing it off as their own work without crediting the source. Image plagiarism is based on image processing, which helps to manipulate and perform operations on image to detect plagiarism. Previously lot of work is done to detect plagiarism on text, but there is no much work done in this area. In this paper an attempt is made to detect difference between the images using image subtraction. The system is also overcomes the vulnerability of re-sizing, compression and color differentiation. The similarity and the difference between the images is displayed using histogram.

Index Terms*: *FMM, Plagiarism, Canny edge detection, CBIR.

I. INTRODUCTION

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.

II. LITERATURE SURVEY

[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 grayscale on which the experiments are performed on all 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.

Revised Manuscript Received on June 7, 2019

Akshay S, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.

Chaitanya B N, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.

Rishabh Kumar, Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Mysuru, India.

The process of retrieving the images from a database according to the visual contents is known as CBIR.

It's nothing but retrieving of images which have similar content of shape, colors and texture etc. This describes about color feature extraction and shape feature extraction. In color feature extraction it uses HSV color space for comparison. Shape feature extraction plays a important role in detecting the actual shape and representation. [8]Detecting plagiarism in images describes about the advance technology of finding the plagiarism in images, as its already implemented for diagrams and flowcharts research based on images is done less. This paper focus on CBIR(Content based image retrieval) for feature extraction such as color, shape and texture. In this paper preprocessing is done in 3 steps: 1. Grey scaling 2. Thresholding 3. Boundary detection and cropping. In grayscaling process RGB values are extracted and calculated using the formula(R+G+B)/2.Thresholdingis a process where a threshold value is calculated and according to that image will be converted into black and white image. After thresholding boundary of the image is calculated and cropping is done according to the requirements.

III. PROPOSED WORK

The proposed work mainly focus 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.

A. Preprocessing

In this phase we are extracting the images from the paper using third-party software and stored in an folder. Pre-processing includes

1. Compression
2. Re-sizing
3. Binarization

The main advantage of image compression is to minimize the original size of the image to lower size. Now a days the image size are very huge so to reduce the computational time FMM algorithm is used. The FMM (Five Modulus Method) algorithm converts an random image into 8*8 block matrix. The block matrix is divided by 5 which reduces the size of the image. We have used FMM for compression as the intended data which is required is not lost in this method. It also reduces the size of the image from higher intensity to lower intensity which helps to compare images more rapidly. FMM firstly converts the input image into gray scaled image and then compresses using threshold value.

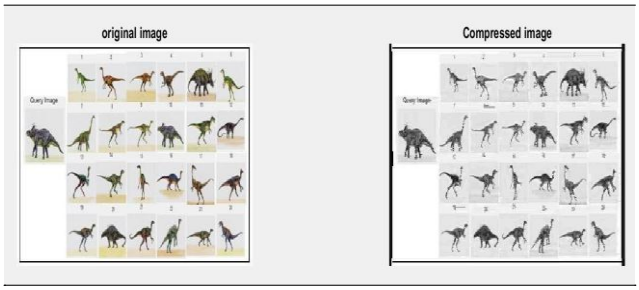

Figure1: Image Compression

To overcome the vulnerability of images with different size, we are re-sizing the images to a fixed size. To get the preferred size of the image the re-sizing is done. The pixel information is changed when the images are re-sized to lower size or upper size. The value of pixel gets added when the images are enlarged.

Figure2:Image Re-sizing

Binarization is the process of converting images into black-0 and white-1 as pixel values. The pixels of the image are replaced by 1 and 0 as per the threshold value. Comparison of the images is done pixel by pixel, the difference determines the similarity between the images. Binary images also helps in reducing the memory storage than normal images. It is also used to overcome the weakness of images with coloured and non coloured.

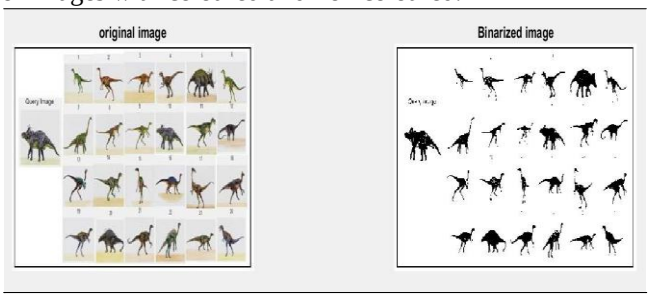

Figure 3: Image Binarization

We also use image subtraction method to find out the difference between the pixel. So the images are converted into binary values. The difference between the images specifies whether the images are same or not. Each pixel value of a image is subtracted with the pixel value of other image.

IV. PROPOSED SYSTEM

In the initial stage we extract images from PDF and store it in a folder.

A. 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.

B. Flow Diagram

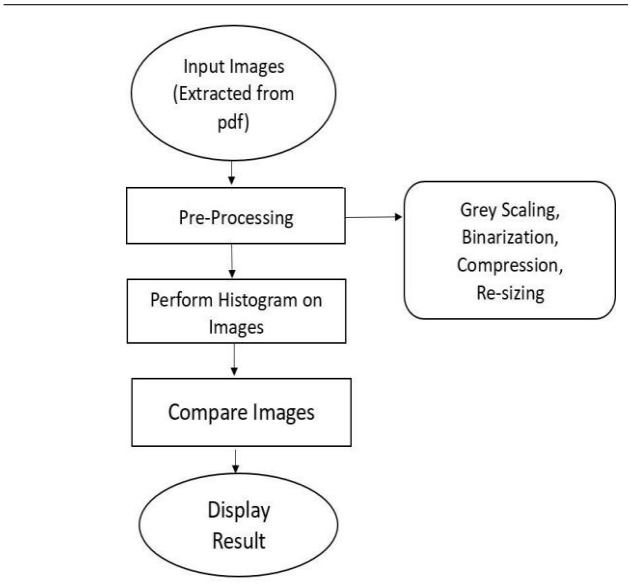

Figure4: Flow Diagram

V. EXPERIMENTAL RESULTS

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.

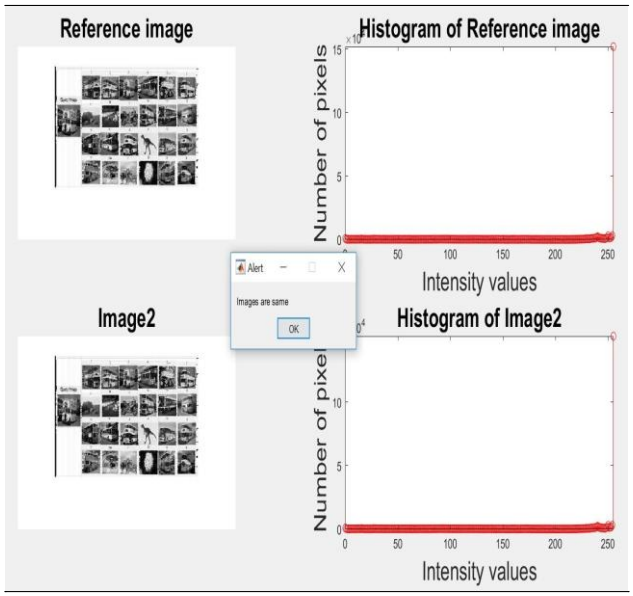

Figure5: Plagiarized Image

The histogram in the above figure represents about the similarity between the images, and it gives a alert box that the images are same, which may leads to plagiarism only if the proper credit to the author is not given.

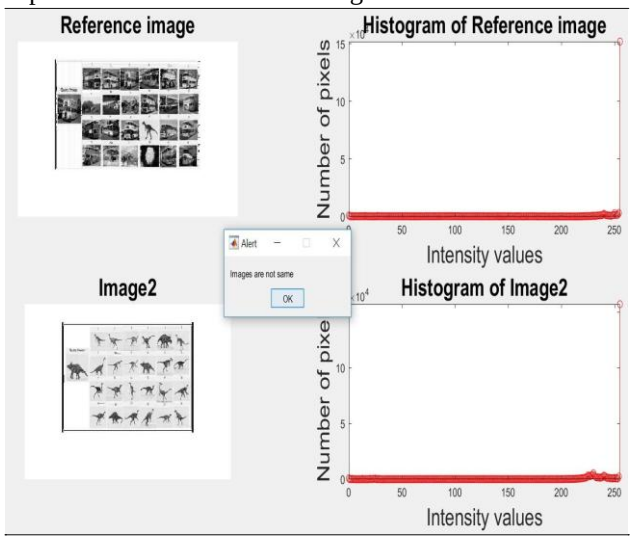

Figure 6: Not-plagiarized Image

Table I: F-Measure Calculation

DATASET	F-MEASURE
Dataset-1	70%

VI. CONCLUSIONS

In this system, We have taken a data set of 80 images which are extracted from research papers through a third party website pdfaid.com. The results obtained through comparison are quite accurate. The F measure value obtained is 70%. The drawback of this system is that, it cannot detect plagiarism on images which are cropped till vital content. The future work on image plagiarism can be implemented on detection of plagiarism of cropped images and re-sized images using advanced techniques. Extraction of images directly from PDF can be done rather than manual technique.

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.

AUTHORS PROFILE

AKSHAY Sreceived M.S degree in Computer Science from University of Mysore in 2012 and is currently pursuing PhD. He is working as an Assistant Professor in Department of Computer Science, Amrita Vishwa Vidyapeetham, Mysore campus. His areas of interest are Pattern Recognition, Digital Signal Processing and Image Processing. He has published many papers in National and International Conferences.

CHAITANYA B N,BCA student at Amrita Vishwa Vidyapeetham, Mysore campus. His area of interest is Image Processing.

RISHABH KUMAR,BCA student at Amrita Vishwa Vidyapeetham, Mysore campus.His area of interest is Image Processing.